UNIVERSITÀ
di VERONA

Department
of ECONOMICS

Working Paper Series
Department of Economics
University of Verona

The Joint Estimate of Singleton and Longitudinal Observations: a GMM Approach for Improved Efficiency

Randolph Luca Bruno, Laura Magazzini, Marco Stampini

WP Number: 4 May 2018

ISSN: 2036-2919 (paper), 2036-4679 (online)

The Joint Estimate of Singleton and Longitudinal Observations: a GMM Approach for Improved Efficiency. [*]

Randolph Luca Bruno[†], Laura Magazzini[‡], Marco Stampini [§]

This version: March 31, 2018

Abstract

We devise an innovative methodology that allows exploiting information from singleton and longitudinal observations for the estimation of fixed effects panel data models. The approach can be applied to join cross-sectional data and longitudinal data, in order to increase estimation efficiency, while properly tackling the potential bias due to unobserved individual characteristics. Estimation is framed within the GMM context and we assess its properties by means of Monte Carlo simulations. The method is applied to an unbalanced panel of firm data to estimate a Total Factor Productivity regression based on the renown Business Environment and Enterprise Performance Survey (BEEPs) database. Under the assumption that the relationship between observed and unobserved characteristics is homogeneous across singleton and longitudinal observations (or across different samples), information from longitudinal data is used to "clean" the bias in the unpaired sample of singletons. This reduces the standard errors of the estimation (in our application, by approximately 8-9 percent) and has the potential to increase the significance of the coefficients.

JEL Classification Numbers: C23, C33, C51.

Keywords: Panel Data, Efficient Estimation, Unobserved Heterogeneity, GMM.

[*] We would like to thank Giorgio Calzolari for very useful comments on a previous version of this paper. Shortcomings are our own.

[†] Randolph Luca Bruno, School of Slavonic and East European Studies, University College London, Gower Street, London, WC1E 6BT, UK and Institute for the Study of Labor (IZA), E-mail: randolph.bruno@ucl.ac.uk.

[‡] Laura Magazzini, Department of Economics, University of Verona, Via Cantarane 24, 37129 Verona, Italy, E-mail: laura.magazzini@univr.it.

[§] Marco Stampini, Social Protection and Health Division, Inter-American Development Bank, 1300 New York Avenue NW, Washington DC, 20577, USA, E-mail: mstampini@iadb.org. The content and findings of this paper reflect the opinions of the authors and not those of the IDB, its Board of Directors or the countries they represent.

1 Introduction

The availability of high quality longitudinal data sets is rare in many field of economic studies. In most cases, researchers rely on repeated, sometimes yearly, cross-sectional data sets. In few others, scholars can also access short panels of data and therefore they routinely use these sources of data to estimate the causal relationship between individual's, household's, firm's characteristics and, for example, the level of wages, welfare, production's efficiency, etc. These are in nature micro-data empirical analyses. Examples include the estimation of the return to education in terms of individual earnings, or of the relationship between the adoption of a technology and firm productivity.

Estimates from cross-sectional data (singletons) are potentially biased because they do not properly account for the relationship between observed time-variant and time-invariant unobserved characteristics. In contrast, fixed effect panel estimates are unbiased, but may be imprecise due to the small size of most longitudinal data sets (e.g. few rounds or few observations per round). As a result, coefficients may not be statistically significant due to low number of degrees of freedom and, as a consequence, low statistical power.

In this paper, following Bruno and Stampini (2009), we show how the singleton observations can be exploited to improve the efficiency of panel estimates based on longitudinal data. In other words, singletons can come from either unpaired observations in panel data sets (e.g., firms that were surveyed only once) or cross-sectional data. Practically all panel data sets contain singletons, due to attrition during data collection. Cross-sectional data is also generally available as a complementary source of information.

The procedure entails two steps. First, information from panel model estimation is used to "clean" the singletons from the bias that comes from the omission of time-invariant unobservable characteristics. Second, the "cleaned" singletons are used jointly with longitudinal observations to produce more precise and *still unbiased* estimates. The gain in efficiency comes from exploiting a higher number of observations.

The validity of the results depends on two key assumptions. Firstly, there must be no time varying unobserved characteristics correlated with either observed characteristics or fixed effects (i.e., the strict exogeneity assumption holds). This is a common assumption of the fixed effects estimation, usually not tested in empirical analysis. The second assumption - which can be tested - is that the relationship between observed and unobserved characteristics is *homogeneous* in the longitudinal observations and in the singletons. We discuss how to test this assumption, and its implication for the applicability of the proposed methodology.

Following the intuition in Bruno and Stampini (2009) and building on it, we devise an innovative procedure within the Generalized Methods-of-Moments (GMM henceforth, see Hansen, 1982) framework.[1] We assess the properties of the methodology by means of Monte Carlo simulations.

The remainder of the paper is organised as follows: in section 2 we identify the conditions under which the methodology can be applied. Section 3 develops a Monte Carlo experiment to assess the efficiency gains. In section 4 we apply the methodology, achieving efficiency gains, to data from fourth and fifth waves of the Business Environment and Economic Performance survey (BEEPs). Section 5 concludes.

2 Exploiting singleton observations in fixed effect estimation

We consider the static panel data model ($i = 1, ..., N; t = 1, ..., T$):

$$y_{it} = x'_{it}\beta + \alpha_i + e_{it} \tag{1}$$

with x_{it} a $k \times 1$ vector of observable characteristics, β a $k \times 1$ vector of parameters to be estimated,[2] α_i the individual fixed effect, and e_{it} an idiosyncratic error term.

The variables in x_{it} are allowed to be arbitrarily correlated with α_i, but not with e_{is} at any time $s = 1, ..., T$, so that the strict exogeneity assumption is satisfied.

Let us denote the singleton observations with the subscript s, that is we denote y_{it} the observations on the units for which we have repeated observations and y_{st} the units for which we only have one observation at time t (unit s is not observed at time periods $r \neq t$, with $r = 1, ..., T$). The observations on the independent variables are defined in an analogous way, as x_{it} for the "panel" units and x_{st} for the singleton observations.

Model (1) is customarily estimated using the within group transformation and the fixed effect estimator (see e.g. Wooldridge, 2004, pag. 265). Under the assumption of strict exogeneity, the fixed effect estimator $\hat{\beta}_{fe}$ is consistent and $\hat{\beta}_{fe} \xrightarrow{p} \beta$. On the contrary, in the case of correlation between x_{it} and α_i, the OLS estimator $\hat{\beta}_{ls}$ is biased and $\hat{\beta}_{ls} \xrightarrow{p} \tilde{\beta}$, with $\tilde{\beta} \neq \beta$. Let us define the OLS bias as $b = \tilde{\beta} - \beta$.

Observations on singleton units are lost in a fixed effect framework, as the within group transformation for those units is identically equal to *zero*. However, we propose

[1] The estimator is easily obtained using available econometric software. A STATA code to implement the proposed procedure is available from the authors upon request.

[2] A constant term is included in β, and x_{it} is defined accordingly.

that the observations on the singleton can be employed to enhance the efficiency and significance of fixed effect estimator, under the assumption that the bias of the OLS estimator does not change across the two samples (longitudinal and singleton data).

In other words, our methodology relies upon the assumption that the correlation between the independent variables and the time-invariant unobservables is constant across samples: the relationship between observable and unobservable characteristics must be homogeneous in the sub-sample of unpaired observations (singletons) and in the balanced panel (longitudinal observations). We call this the *homogeneity hypothesis*.[3] We need that the *homogeneity hypothesis* and the common variance between panel and cross sections are both satisfied.

The estimation method we propose is framed within a GMM approach.

First, consider the fixed effect estimator on the longitudinal observations.[4] This can be obtained as the OLS estimator of the within-group transformed data, thus relying on the following moment condition:

$$E[\ddot{x}_{it}(\ddot{y}_{it} - \ddot{x}'_{it}\beta)] = 0 \tag{2}$$

with $\ddot{z}_{it} = z_{it} - \bar{z}_i$ with $\bar{z}_i = \sum_t z_{it}/T$ $(z = y, x)$.

The fixed effect panel data estimator has also an instrumental variable interpretation (Verbeek, 2008, pag. 354), in which the regression model in (1) is considered, and each explanatory variable is instrumented by its value in deviation from the individual specific mean $\ddot{x}_{it}$.[5] As a result, the following moment condition is exploited leading to the fixed effect estimate of β:[6]

$$E[\ddot{x}_{it}(y_{it} - x'_{it}\beta)] = 0 \tag{3}$$

Now consider the OLS estimator. As already mentioned, in the case of correlation between the explanatory variables and the individual effect the OLS estimator is biased. Still, given the properties of the OLS estimator, the following condition holds:

$$E[x_{it}(y_{it} - x'_{it}(\beta + b))] = 0 \tag{4}$$

A GMM estimator based on the moment conditions in (3) and (4) will produce the

[3] See also Bruno and Stampini (2009) for a full formalisation.

[4] Singletons can be included in the estimation, but their presence will not change the results as $\ddot{x}_{st} = 0$ and $\ddot{y}_{st} = 0$ for each s and t.

[5] When the model contains a constant term, the moment condition $E[y_{it} - x'_{it}\beta] = 0$ is also considered. That is, the within group transformation is not considered for the constant term.

[6] Again, singleton observations will *not* contribute to estimation.

fixed effect estimator of β. As we are adding k moment conditions and k parameters (the OLS bias of each coefficient in b), the estimation of β is unaffected by the new moment condition (4).

However, the proposed set up allows us to exploit the additional information provided by the availability of singleton observations (y_{st}, x_{st}). Under the assumption of homogeneity, the following moment condition can also be exploited, leading to additional information for the estimation of β (and b):

$$E[x_{st}(y_{st} - x'_{st}(\beta + b))] = 0 \tag{5}$$

In the following section a set of Monte Carlo experiments is presented to show that the proposed methodology can indeed increase the efficiency of the fixed effect estimator.

3 Monte Carlo experiments

First, let us consider the longitudinal data as:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + e_{it} \tag{6}$$

where $\beta_0 = 0$, $\beta_1 = 1$, $\alpha_i \sim N(0, \sigma^2_\alpha)$, and $e_{it} \sim N(0, \sigma^2_e)$, with $\sigma^2_e = 1$, $\sigma^2_\alpha = 0.25, 1, 4$, $i = 1, ..., N_p$ (with $N_p = 100$), and $t = 1, ..., T$ (we will consider $T = 2, 4$). As for the independent variable, we let:

$$x_{it} = \theta\, \alpha_i + w_{it}$$

with $w_{it} \sim N(0, \sigma^2_w)$, $\theta = 0.2, 0.8$, and $\sigma^2_w = 0.8, 1.2$.

In order to generate the singleton observations, we consider the same data generating process as the longitudinal observations, but we only generate observations at time $t = 1$:[7]

$$y_{s1} = \beta_0 + \beta_1 x_{s1} + \alpha_s + e_{st}$$

so that the homogeneity assumption is satisfied. The sample size of the singleton data is set as $N_s = s \times N_p$. The total number of observations is therefore $N = N_p + N_s = N_p(1 + s)$.

Results of Monte Carlo experiments are reported in Table 1 (based on 10,000 replications).

[7] The first time period is considered for simplicity. Singleton observations can be observed in any time periods. Indeed, in the empirical application presented in the next Section, singletons are distributed across all time periods, as usually common in micro-databases.

T	σ_w^2	θ	σ_α^2	FE	GMM $s = 0.5$	GMM $s = 1$	GMM $s = 1.5$
2	0.8	0.2	0.25	1.000 (.1137)	1.004 (.1104)	1.001 (.1069)	1.001 (.1029)
2	0.8	0.2	1.00	1.000 (.1137)	1.001 (.1117)	1.002 (.1099)	1.002 (.1071)
2	0.8	0.2	4.00	1.000 (.1137)	1.001 (.1131)	1.002 (.1128)	1.002 (.1112)
2	0.8	0.8	0.25	1.000 (.1137)	1.001 (.1107)	1.002 (.1079)	1.002 (.1043)
2	0.8	0.8	1.00	1.000 (.1137)	1.001 (.1123)	1.002 (.1111)	1.003 (.1089)
2	0.8	0.8	4.00	1.000 (.1137)	1.001 (.1133)	1.003 (.1131)	1.002 (.1119)
2	1.2	0.2	0.25	1.000 (.0928)	1.000 (.0901)	1.001 (.0828)	1.001 (.0840)
2	1.2	0.2	1.00	1.000 (.0928)	1.000 (.0913)	1.001 (.0897)	1.001 (.0874)
2	1.2	0.2	4.00	1.001 (.0928)	1.000 (.0923)	1.002 (.0921)	1.002 (.0907)
2	1.2	0.8	0.25	1.000 (.0928)	1.000 (.0903)	1.001 (.0878)	1.002 (.0848)
2	1.2	0.8	1.00	1.000 (.0928)	1.001 (.0915)	1.002 (.0904)	1.002 (.0886)
2	1.2	0.8	4.00	1.000 (.0928)	1.001 (.0925)	1.002 (.0923)	1.002 (.0913)
4	0.8	0.2	1.00	1.000 (.0645)	1.000 (.0641)	1.001 (.0633)	1.000 (.0620)
4	0.8	0.8	1.00	1.000 (.0645)	1.000 (.0643)	1.001 (.0639)	1.001 (.0630)
4	1.2	0.2	1.00	1.000 (.0529)	1.000 (.0523)	1.001 (.0517)	1.000 (.0505)
4	1.2	0.8	1.00	1.000 (.0528)	1.001 (.0525)	1.001 (.0521)	1.001 (.0513)

Table 1: Results of Monte Carlo simulations, mean and standard deviations (in parenthesis) of estimated β_1, 10,000 replications

Overall, the proposed methodology allows to increase the efficiency of the parameters' estimate, without affecting the property of consistency, though. In some instances, a 10% decrease in the standard errors is recorded. As expected, efficiency gains are increasing in s, i.e. in the number of singletons. Furthermore, the efficiency gains depend on the variance of the individual effects, and they are more pronounced in the case of smaller σ_α^2 and smaller θ.

4 An empirical application

We apply the proposed methodology to estimate a total factor productivity (TFP) regression on data from the European Bank of Reconstruction and Development – World Bank Business Environment and Enterprise Performance Survey (BEEPS).[8] The survey contains firm-level data on broad range of variables about the business environment and performance of firms. We considered data from fourth (IV) and fifth (V) waves, from the years 2007 and 2011-2012 respectively.[9]

[8] See http://ebrd-beeps.com and https://openknowledge.worldbank.org/handle/10986/9393.

[9] The following countries are included in the survey: Albania, Armenia, Azerbaijan, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Estonia, FYR Macedonia, Georgia, Hungary, Kaza-

	No missing		No outliers	
	N	%	N	%
BEEPS IV only	2,475	52.44	2,031	54.28
BEEPS V only	1,974	41.82	1,532	40.94
Both BEEPS IV and V	271	5.74	179	4.78
Total	4,720	100.0	3,742	100.0

Table 2: Structure of the data

Following a well established literature on the estimation of the TFP (Hsieh and Klenow (2009); Syverson (2011)), and the related literature applying this estimation to BEEPs data (Hellman, Jones and Kaufmann (2003); Kuboniwa (2011)), we estimate the following equation:

$$lnY_{it} = \alpha_i + \beta_1 lnL_{it} + \beta_2 lnK_{it} + e_{it} \tag{7}$$

The data on sales and capital are reported in local currency units. Values have been deflated and converted to US dollars using the consumer price index and the exchange rate provided by the World Bank (source: World Bank, World Development Indicators).[10]

In our dataset, after removing outliers and observations with missing data,[11] we have 358 longitudinal observations (179 units, each one observed twice) and 3,563 singletons (2,031 from the fourth wave, and 1,532 from the fifth) (Table 2).

Before applying the GMM methodology, we need to test whether the homogeneity assumption is satisfied, that is we check that the OLS estimates do not differ in the longitudinal (panel) and singleton samples. This is accomplished by estimating (by OLS) a fully interacted model and testing whether the OLS coefficients differ in the longitudinal and singleton samples. The F-test for the null hypothesis of no difference

khstan, Kosovo, Kyrgyz Republic, Latvia, Lithuania, Moldova, Mongolia, Montenegro, Poland, Romania, Russia, Serbia, Slovak Republic, Slovenia, Tajikistan, Turkey, Ukraine, and Uzbekistan. Data for Cyprus and Greece are also included in BEEPS V. However, differently from most countries in the analysis, data refers to the year 2014. We decided not to include those countries in the analysis. In addition, data on consumer price index are not available for these countries.

[10] Data on exchange rate and consumer price index are not available for Belarus, Uzbekistan, Estonia, Slovak Republic, Slovenia; therefore these countries are excluded from the analysis. We take into account the fiscal year the data refers to, as indicated in the BEEPs data.

[11] For each country, we computed the median (m) of all the variables and the inter-quartile range (iqr), and considered as outliers those observations with values outside the interval defined by $m \pm 1.5iqr$. We removed about 20% observations.

	No missing		No outliers	
Time dummy	no	yes	no	yes
F-test	1.76	1.49	.462	.381
p-value	.152	.202	.709	.823

Table 3: Results of the test of homogeneity

	No missing				No outliers			
Variable	FE	GMM	FE	GMM	FE	GMM	FE	GMM
$\ln L$	.611***	.734***	.554***	.662***	.653***	.650***	.607***	.602***
	(.169)	(.168)	(.162)	(.159)	(.121)	(.116)	(.131)	(.125)
$\ln K$	.151***	.117**	.121**	.084	.156***	.166***	.126**	.132***
	(.057)	(.052)	(.057)	(.053)	(.057)	(.049)	(.053)	(.047)
Time dummy	no	no	yes	yes	no	no	yes	yes
Test CRS	2.10	.770	4.18	2.48	2.37	2.23	3.77	3.93
[p-value]	[.147]	[.380]	[.041]	[.115]	[.124]	[.136]	[.052]	[.047]
J-test	–	5.06	–	5.68	–	1.34	–	1.49
[p-value]		[.168]		[.225]		[.719]		[.829]

Table 4: Results of the econometric estimation; standard errors are robust to clustering across units; p-value of reported test statistics among squared brackets

across the two samples is reported in Table 3.[12] We consider both the specification in (7) and a specification including a time dummy.

The homogeneity assumption is never rejected. We can therefore use the GMM framework we propose to increase the efficiency of fixed effect estimates.

Estimation results are presented in Table 4, that reports both the fixed effect estimator and the GMM estimator implementing the correction methodology. The Table also shows the result of a test for the null hypothesis of constant returns to scale (CRS), i.e. $H_0 : \beta_1 + \beta_2 = 1$, and the J-test for the validity of over-identifying restrictions.

The first visible result is the decrease of the Standard Errors size for each and every column of the GMM estimation compared to the FE. The gain of efficiency is approximately 8-9 percent, i.e. a non-negligible by the standard of this type of regressions analysis in the TFP literature. In only one instance, this increases the level of significance of the coefficient estimation (for the elasticity to capital, in the model with time

[12] Consider the OLS regression for the full sample $y_{jt} = x'_{jt}\beta + w_{jt}$ with $j = i$ for longitudinal observations and $j = s$ for singleton data. Define a dummy variable for the panel observations, that is $d_{jt} = 1$ if $j = i$, 0 otherwise. The fully interacted model is $y_{jt} = x'_{jt}\beta + d_{jt}\, x'_{jt}\delta + w_{jt}$. The test of homogeneity corresponds to testing the null hypothesis $H_0 : \delta = 0$, see also Bruno and Stampini (2009).

dummy). Estimates of the elasticities to labor and capital are in line with the findings of the existing literature, i.e. in the order of 60-70 percent for labour versus 10-20 percent for capital (see Syverson (2011)).

5 Conclusions

In this paper we devise an innovative procedure that exploits singleton (i.e. unpaired or cross-sectional) observations to increase the efficiency of panel data estimates of microeconomic relationships. The estimation procedure we propose is built within a Generalised Methods of Moments framework.

The availability of longitudinal data allows us to properly tackle the the potential bias due to the correlation between variables of interest and unobserved time-invariant individual characteristics. The use of the singletons allows reducing the standard errors of the panel estimates, with the potential to increase their significance.

Our procedure relies on the assumption that the relationship between observed and unobserved characteristics is homogeneous across samples. The assumption can be easily tested using an 'interacted' OLS regression.

We show efficiency gains through a set of Monte Carlo experiments. We then apply our procedure to the estimation of a Total Factor Productivity regression using data from the Business Environment and Enterprise Performance survey. Results show increases in efficiency in all models' specification, in the order of 8-9 percent.

The procedure can be applied to the estimation of other microeconomic relationships, such as that of the returns to education in terms of earnings, production function, economic performance, etc. In general, it can be useful to researchers dealing with small panel data sets in a context of richer (vis-a-vis the panel) availability of cross-sectional data. Similarly, it can be used to recover information from large attrition's units in large panel data sets. The methodology could be implemented to a wide range of contexts and databases.

References

Becketti, S., Gould, W., Lillard, L., Welch, F., (1998). "The panel study of income dynamics after fourteen years: An evaluation", *Journal of Labor Economics*, vol. 6, pp. 472-492.

Breusch, Trevor, Grayham E. Mizon and Peter Schmidt (1989). "Efficient Estimates Using Panel Data", *Econometrica*, vol. 57(3), pp. 695-700.

Browning, Martin, Angus Deaton and Margaret Irish (1985)."A Profitable Approach to Labour Supply and Commodity Demands over the Life.Cycle", *Econometrica*, vol. 53(3), pp. 503-543.

Bruno, R.L., and Stampini, M., 2009: Joining Panel Data with Cross-Sections for Efficiency Gains, *Giornale degli Economisti e Annali di Economia, Nuova Serie*, **68**(2), 149-173.

Deaton, Angus (1985). "Panel Data from Time Series of Cross-Sections", *Journal of Econometrics*, vol. 30, pp. 109-126.

Hansen, L.: 1982, Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica* **50**, 1029-1054.

Hausman, J. A. and D. A. Wise (1979). "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment", *Econometrica*, vol. 47, pp. 455-473.

Hellman, JS , G Jones and D Kaufmann (2003). "Seize the state, seize the day: state capture and influence in transition economies", *Journal of comparative economics*

Hirano, Keisuke, Guido W. Imbens, Geert Ridder and Donald B. Rubin (2001). "Combining Panel Data Sets with Attrition and Refreshment Samples", *Econometrica*, vol. 69(6), pp. 1645-1659.

Hsieh, CT and PJ Klenow (2009). "Misallocation and manufacturing TFP in China and India", *The Quarterly journal of economics*

Kuboniwa, M (2011). "The Russian growth path and TFP changes in light of estimation of the production function using quarterly data", *Post-Communist Economies.*

Moulton, B. R. (1986). "Random group effects and the precision of regression estimates", *Journal of Econometrics*, vol. 32, pp. 385-397.

Nijman, Theo and Marno Verbeek (1990). "Estimation of Time-Dependent Parameters in Linear Models Using Cross-Sections, Panel, or Both", *Journal of Econometrics*, vol. 46, pp. 333-346.

Pitt, Mark M., Mark R. Rosenzweig and Donna M. Gibbons (1993), "The Determinants and consequences of the Placement of Goverment Programs in Indonesia", *The World Bank Economic Review*, vol. 7, pp. 319-348.

Ridder, Geert (1992), "An empirical evaluation of some models for non-random attrition in panel data" *Structural Change and Economic Dynamics*, vol. 3(2), pp. 337-355.

Ridder, Geert and Robert Moffit (2007), "The Econometrics of Data Combination" in *Handbook of Econometrics*, ed. by James J. Heackman - Edward E. Leamer, vol. 6(part 2), pp. 5469-5547.

Stampini, M. and B. Davis (2006). "Discerning Transient From Chronic Poverty In Nicaragua: Measurement With A Two Period Panel Data Set". *European Journal of Development Research*, vol. 18(1), pp. 105-130.

Syverson, C (2011). "What determines productivity?". *Journal of Economic literature.*

Verbeek, Marno and Theo Nijman (1992). "Can Cohort Data be Treated as Genuine Panel Data?", *Empirical Economics*, vol. 17, pp. 9-23.

Verbeek, M., 2008: *A guide to modern econometrics*, John Wiley & Sons.

Wooldridge, J.M., 2004: *Econometrics of Cross Section and Panel Data*, The MIT Press.