# Exploiting Information from Singletons in Panel Data Analysis: a GMM Approach

**Abstract**

We propose a novel procedure, built within a Generalized Method of Moments framework, which exploits unpaired observations (singletons) to increase the efficiency of longitudinal fixed effect estimates. The approach allows increasing estimation efficiency, while properly tackling the bias due to unobserved time-invariant characteristics. We assess its properties by means of Monte Carlo simulations, and apply it to a traditional Total Factor Productivity regression, showing efficiency gains of approximately 8-9 percent.

JEL Classification Numbers: C23, C33, C51.

*Keywords:* Singletons, Panel Data, Efficient Estimation, Unobserved Heterogeneity, GMM

## 1. Introduction

Within a static panel data framework, fixed effect (FE) estimation allows for correlation between explanatory variables and unobserved individual components. In case of spherical idiosyncratic error term, the FE estimator is efficient. However, by imposing additional assumptions on the model structure, more efficient estimators can be developed. For example, Hausman and Taylor [6], Amemiya and MaCurdy [1], Breusch *et al.* [2] consider the case in which some of the variables in the model are uncorrelated with the individual effects.

We consider a different approach that achieves higher efficiency by exploiting information from singletons, i.e. sample units observed only once. These units are ignored in the FE estimation, as their within group (WG) transformation equals zero. Our innovative procedure is built within the Generalized Method of Moments (GMM) framework [5].

We build on Bruno and Stampini [3], whose three-step methodology combines longitudinal data and unpaired cross-sectional observations. Information from the former is used to correct the bias of the latter which characterizes OLS estimates in the presence of (time-invariant) omitted variables. Being based on a larger set of observations with respect to the standard FE

approach, the procedure leads to efficiency gains. We extend this methodology along two dimensions. First, estimation is framed within GMM. Second, while Bruno and Stampini [3] used singletons from different cross sectional datasets, we apply the procedure to a single panel dataset, where the singletons are unpaired units observed only once.

The validity of the results depends on the plausible assumption that the relationship between observed and unobserved characteristics is *homogeneous* in longitudinal observations *and* singletons.

The paper is organised as follows. In Section 2 we identify the conditions under which the methodology can be applied, and discuss how to test the validity of the underlying assumption. Section 3 develops a Monte Carlo experiment to assess the efficiency gains. In Section 4 we apply the methodology to data from two waves of the Business Environment and Economic Performance survey (https://ebrd-beeps.com/). Section 5 concludes.

## 2. Exploiting singleton observations in FE estimation

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

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

with $x_{it}$ a $k \times 1$ vector of observable characteristics, $\beta$ a $k \times 1$ vector of parameters to be estimated,[1] $\alpha_i$ the individual fixed effect, and $e_{it}$ an idiosyncratic error term. We allow the panel to be unbalanced as $T_i$ observations are available on unit $i$, with $T_i = 1, ..., T$, and $T > 1$.[2] The variables in $x_{it}$ are allowed to be arbitrarily correlated with $\alpha_i$, but not with $e_{is}$ at any time ($s = 1, ..., T_i$), so that the strict exogeneity assumption is satisfied. Let us denote the singletons with the subscript $i = s$, so that $T_s = 1$. Also denote $N_p$, the number of units observed more than once (units $i$ with $T_i > 1$), and $N_s$ the number of singletons ($N = N_s + N_p$). The total number of observations is $n = N_s + \sum_{i=1}^{N_p} T_i$.

The FE estimator can be obtained by using the WG transformation, in which individual means are computed over the observed time period (see,

[1] A constant term can be included in $\beta$, with $x_{it}$ defined accordingly.

[2] If all sample units were observed only once, a cross-sectional dataset would be available, and the FE estimator would not be defined. The method we propose can be applied to datasets in which a sufficient number of longitudinal observations is available for application of the FE estimator.

e.g. Verbeek, 2008, pag. 382):

$$\hat{\beta}_{fe} = \left( \sum_{i=1}^{N} \sum_{t=1}^{T_i} \ddot{x}_{it} \ddot{x}'_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T_i} \ddot{x}_{it} \ddot{y}_{it}$$

where $\ddot{z}_{it} = z_{it} - \bar{z}_i$ denotes the WG transformation with $\bar{z}_i = \sum_t z_{it}/T_i$ $(z = y, x)$. As the WG transformation is identically zero for the singletons, only the units observed more than once are employed in estimation:

$$\hat{\beta}_{fe} = \left( \sum_{i=1}^{N_p} \sum_{t=1}^{T_i} \ddot{x}_{it} \ddot{x}'_{it} \right)^{-1} \sum_{i=1}^{N_p} \sum_{t=1}^{T_i} \ddot{x}_{it} \ddot{y}_{it}$$

Consistency of the FE estimator relies on the strict exogeneity assumption, and, in the case of unbalanced panel datasets, on assumptions about the process driving attrition. When observations are missing at random [7, pag. 381],[3] the following holds:

$$\text{plim}_{N_p \to \infty} \hat{\beta}_{fe} = \beta.$$

Note that, consistency of the FE estimator relies on $N_p \to \infty$.

The FE estimator has also an instrumental variable interpretation [7, pag. 354] in which the regression model in (1) is considered, and each explanatory variable is instrumented by its deviation from the individual specific mean. We can therefore write:

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

In contrast, in case of correlation between the explanatory variables $x_{it}$ and the FE $\alpha_i$, the OLS estimator $\hat{\beta}_{ls}$ is biased and

$$\text{plim}_{N \to \infty} \hat{\beta}_{ls} = \tilde{\beta}$$

with $\tilde{\beta} \neq \beta$.[4] We denote the OLS bias as $b = \tilde{\beta} - \beta$. Still, we can write the following moment conditions:

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

[3] More general assumptions can also be considered: see [7, pag. 383].

[4] Asymptotic properties of the OLS estimator would be obtained for $n \to \infty$. As we consider a panel data framework, we let $N$ grow large with fixed $T$. This is accomplished when $N_p \to \infty$, even for fixed $N_s$. As a result, OLS allows exploiting information from all the units in the sample.

As we are adding $k$ moment conditions and $k$ parameters (the OLS bias of each coefficient in $b$), a GMM estimator based on the moment conditions in (2) and (3) will produce the FE estimator of $\beta$.

However, under the assumption that the OLS bias is *homogeneous* in the two sub-samples of longitudinal and singleton observations, the following additional moment conditions can be exploited:

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

We call this assumption the *homogeneity hypothesis* (see also Bruno and Stampini, 2009). A consistent estimate of $\beta$, with increased efficiency, can be obtained using a GMM approach that exploits the moment conditions (2), (3) *and* (4). We expect the homogeneity assumption to hold when the observations are missing at random.

In our setting, the OLS bias can be ascribed to the presence of correlation between the individual heterogeneity and the regressors:

$$b = \left( \sum_{i=1}^{N} \sum_{t=1}^{T_i} x_{it} x'_{it} \right)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T_i} x_{it} \alpha_{it}$$

with probability limit equal to the ratio between $\text{cov}(x_{it}, \alpha_i)$ and the variance of $x_{it}$. The homogeneity assumption therefore requires that the covariance between $x$ and $\alpha$ and the variance of $x$ are the same in the full sample and among the singletons.

To better understand this assumption, consider a data generating process in which $y^*_{it} = x^{*\prime}_{it}\beta + \alpha_i + e_{it}$, and introduce random missing observations, so that for some (randomly-selected) units $y^*_{it}$ and $x^*_{it}$ are observed only once (at a randomly selected time period). The homogeneity assumption is satisfied when, for example, $\text{cov}(x_{it}, \alpha_i)$ and $\text{var}(x_{it})$ are constant over time. It is also satisfied under more general data generating processes, as the random selection hypothesis implies that the distribution of $y$ conditional on $x$ for the singletons is the same as the distribution of $y^*$ conditional on $x^*$ [7, pag. 381].

The homogeneity assumption can be easily tested through a fully interacted OLS regression model. Define a dummy variable for the singleton observations, that is $d_{it} = 1$ if $i = s$, 0 otherwise. The fully interacted model is $y_{it} = x'_{it}\beta + d_{it}\, x'_{it}\delta + w_{it}$. The test of homogeneity corresponds to testing the null hypothesis $H_0 : \delta = 0$. Being an overidentified model, the Hansen test can also be considered to check the validity of the underlying assumptions.

In the following section, a set of Monte Carlo experiments shows that the proposed methodology can indeed increase efficiency relative to the FE

estimator. This is not surprising, as we are exploiting additional moment conditions within a GMM framework. It is important to stress that the underlying assumptions are likely to be satisfied in most empirical applications, as they simply require the singleton data to be produced by the same data generating process as the longitudinal observations.

## 3. Monte Carlo experiments

We consider the following data generating process:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \sigma_\alpha \alpha_i + \sigma_e e_{it} \tag{5}$$

where $\beta_0 = 0$, $\beta_1 = 1$, $\alpha_i \sim N(0,1)$, and $e_{it} \sim N(0,1)$, $i = 1, ..., N$, and $t = 1, ..., T$. We retain all generated $T$ values for the $N_p$ longitudinal observations, whereas only one observation at a randomly selected time period is considered for the $N_s$ singletons. The total number of observations is therefore $n = T\,N_p + N_s$. Let $\lambda = N_s/N$ denote the share of singleton units.

As for the independent variable, we let:

$$x_{it} = \delta_1 \alpha_i + \delta_2 \gamma_i + \delta_3 w_{it}$$

with $\gamma_i$ and $w_{it}$ normally distributed with mean zero and variance one.

The data generating process satisfies the homogeneity assumption.

The FE estimator only exploits the within variance of $x$, $\delta_3^2$. Changes in the parameters $\delta_1$ and $\delta_2$ only affect the between variability of $x_{it}$ (equal to $\delta_1^2 + \delta_2^2$). Therefore, changes in $\delta_1$ and $\delta_2$ do not alter the variance of the FE estimator, but influence the variance of the (potentially biased) OLS estimator and of our GMM approach. Note also that the parameter $\delta_1$ drives the correlation between $x_{it}$ and the composite error term in the data generating process of $y_{it}$ ($\sigma_\alpha \alpha_i + \sigma_e e_{it}$), therefore affecting the bias of the OLS estimator. As for the composite error terms, the variance of the FE estimator will be only affected by changes in $\sigma_e$, whereas changes in $\sigma_\alpha$ would also influence the variance of our GMM estimator (and of the OLS estimator).

We also consider a set of experiments in which we fix $\delta_2 = 0$, and allow $\delta_1$ and $\delta_3$ to vary over time. In particular, we consider $\delta_1 = 2$ for $t \leq T/2$, $\delta_1 = 0$ for $t > T/2$; and $\delta_3 = 0.5$ for $t \leq T/2$, $\delta_3 = 1.5$ for $t > T/2$. In this way, both $\text{cov}(x_{it}, \alpha_i)$ and $\text{var}(x_{it})$ change over time.

Results of Monte Carlo experiments based on 10,000 replications are reported in Table 1. The table shows the FE estimator (WG transformation) and GMM estimates of the parameter of interest $\beta_1$ (a two-step approach is

| | | | | | | | FE | | GMM | | efficiency |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $T$ | $\lambda$ | $\delta_1$ | $\delta_2$ | $\delta_3$ | $\sigma_\alpha^2$ | $\sigma_e^2$ | mean | s.dev. | mean | s.dev. | gain (var.) |
| 2 | 0.5 | .707 | .707 | 1 | 1 | 1 | .9996 | .1006 | 1.001 | .0988 | 3.47% |
| 2 | 0.5 | 1 | 0 | 1 | 1 | 1 | .9996 | .1006 | 1.001 | .0984 | 4.34% |
| 2 | 0.5 | 0 | 1 | 1 | 1 | 1 | .9996 | .1006 | .9996 | .0990 | 3.16% |
| 2 | 0.5 | 1 | 1 | 1 | 1 | 1 | .9996 | .1006 | 1.000 | .0996 | 2.08% |
| 2 | 0.5 | .707 | .707 | 2 | 1 | 1 | .9998 | .0503 | 1.000 | .0488 | 5.76% |
| 2 | 0.5 | .707 | .707 | 1 | 2 | 1 | .9996 | .1006 | 1.001 | .0996 | 2.04% |
| 2 | 0.5 | .707 | .707 | 1 | 1 | 2 | .9994 | .1423 | 1.001 | .1388 | 4.86% |
| 2 | 0.9 | .707 | .707 | 1 | 1 | 1 | 1.000 | .1009 | 1.002 | .0968 | 7.58% |
| 2 | 0.9 | 1 | 0 | 1 | 1 | 1 | 1.000 | .1009 | 1.004 | .0956 | 10.11% |
| 2 | 0.9 | 0 | 1 | 1 | 1 | 1 | 1.000 | .1009 | 1.000 | .0976 | 6.42% |
| 2 | 0.9 | 1 | 1 | 1 | 1 | 1 | 1.000 | .1009 | 1.002 | .0983 | 5.02% |
| 2 | 0.9 | .707 | .707 | 2 | 1 | 1 | 1.000 | .0504 | 1.001 | .0470 | 13.06% |
| 2 | 0.9 | .707 | .707 | 1 | 2 | 1 | 1.000 | .1009 | 1.002 | .0985 | 4.61% |
| 2 | 0.9 | .707 | .707 | 1 | 1 | 2 | 1.000 | .1426 | 1.003 | .1344 | 11.24% |
| 4 | 0.5 | .707 | .707 | 1 | 1 | 1 | .9996 | .0583 | 1.000 | .0573 | 3.43% |
| 4 | 0.9 | .707 | .707 | 1 | 1 | 1 | 1.000 | .0576 | 1.002 | .0551 | 8.41% |
| 10 | 0.5 | .707 | .707 | 1 | 1 | 1 | 1.000 | .0329 | 1.001 | .0327 | 0.94% |
| 10 | 0.9 | .707 | .707 | 1 | 1 | 1 | 1.000 | .0334 | 1.001 | .0327 | 4.28% |
| 20 | 0.5 | .707 | .707 | 1 | 1 | 1 | 1.000 | .0227 | 1.000 | .0227 | 0.37% |
| 20 | 0.9 | .707 | .707 | 1 | 1 | 1 | .9999 | .0231 | 1.001 | .0229 | 1.92% |
| *DGP with $\delta_1$ and $\delta_3$ varying over time* | | | | | | | | | | | |
| 2 | 0.5 | 2;0 | 0 | 1.5;.5 | 1 | 1 | .9988 | .0554 | 1.000 | .0538 | 5.89% |
| 2 | 0.5 | 2;0 | 0 | 1.5;.5 | 2 | 1 | .9988 | .0554 | 1.001 | .0546 | 3.06% |
| 2 | 0.5 | 2;0 | 0 | 1.5;.5 | 1 | 2 | .9982 | .0783 | 1.000 | .0749 | 8.66% |
| 2 | 0.9 | 2;0 | 0 | 1.5;.5 | 1 | 1 | 1.000 | .0561 | 1.000 | .0495 | 22.22% |
| 2 | 0.9 | 2;0 | 0 | 1.5;.5 | 2 | 1 | 1.000 | .0561 | 1.001 | .0518 | 14.73% |
| 2 | 0.9 | 2;0 | 0 | 1.5;.5 | 1 | 2 | 1.000 | .0794 | 1.002 | .0670 | 28.82% |
| 4 | 0.5 | 2;0 | 0 | 1.5;.5 | 1 | 1 | .9994 | .0361 | 1.000 | .0350 | 5.71% |
| 4 | 0.9 | 2;0 | 0 | 1.5;.5 | 1 | 1 | 1.000 | .0359 | 1.002 | .0323 | 19.28% |

Table 1: Results of Monte Carlo simulations, mean and standard deviations (s.dev.) of estimated $\beta_1$, 10,000 replications, $N_p = 100$

considered).[5] $N_p$ is set equal to 100. Efficiency gains, in the last column of the table, are computed on the basis of variances.

Overall, the proposed methodology allows increasing the efficiency of the parameter's estimate, without introducing a bias. As expected, efficiency gains increase with the share of singletons.[6] Larger gains are obtained by increasing the within variability of $x$ ($\delta_3^2$) and the within variability of the error term ($\sigma_e^2$), and reducing the between variability of $x$ (the sum of $\delta_1^2$ and $\delta_2^2$) and the between variability of the error term ($\sigma_\alpha^2$). When the between variability of $x$ is held constant,[7] higher gains are associated with larger correlation between $x$ and $\alpha$ (equal to $\delta_1/\sqrt{\delta_1^2+\delta_2^2+\delta_3^2}$ ). In contrast, efficiency gains are smaller in long panels, characterized by a high number of time periods (see the cases with $T = 10, 20$). In these instances, the singletons provide little additional information.

## 4. An empirical application to the BEEPs data

We apply the proposed methodology to estimate a total factor productivity (TFP) regression on data from the Business Environment and Enterprise Performance Survey (BEEPS).[8] The survey contains firm-level business environment and performance data. We consider data from fourth (IV) and fifth (V) waves, from the years 2007 and 2011-2012 respectively.

Following the literature that estimates the TFP with BEEPs data [4], we estimate the following equation:

$$\ln Y_{it} = \alpha_i + \beta_1 \ln L_{it} + \beta_2 \ln K_{it} + e_{it} \tag{6}$$

After removing outliers[9] and missing data, we have 358 longitudinal ob-

[5]The estimates are obtained using the *gmm* command in STATA. We used the following syntax: *gmm (mc1: y - b1*x -b0) (mc2: y - (b1 +d1)*x -(b0+d0)) (mc3: ys - (b1 +d1)*xs -(b0+d0)) , instruments(mc1: xdemeaned) instruments(mc2: x) instruments(mc3: xs) twostep winitial(unadjusted, independent) nocommonesample vce(cluster id) wmatrix(cluster id)*, where *y* and *x* contain the dependent and independent variables on all observations, *ys* and *xs* contain the singletons (missing for longitudinal observations), and *xdemeaned* contains the WG transformation of the longitudinal units (*xdemeaned*= 0 for the singletons).

[6]In unreported Monte Carlo simulations, we also consider $\lambda = 0.1$. In this instance, due to the limited number of available singletons, the proposed approach does not lead to efficiency gains.

[7]When $\delta_1 = \delta_2 = 0.707$, the sum $\delta_1^2 + \delta_2^2$ is about 1 (0.9997), as in the cases (i) $\delta_1 = 0$, $\delta_2 = 1$ and (ii) $\delta_1 = 1$, $\delta_2 = 0$.

[8]See https://ebrd-beeps.com.

[9]For each country, we computed the median ($m$) of all the variables and the inter-

servations (179 units, each one observed twice) and 3,563 singletons (2,031 from the wave IV, and 1,532 from wave V) so that $\lambda = 0.95$. The data on sales and capital are reported in local currency units. Their values were deflated and converted to US dollars using the consumer price index and the exchange rate provided by World Development Indicators.

FE and two-step GMM estimates are presented in Table 2.[10] We estimate both equation (6) and an alternative specification including time dummies. At the bottom of Table 2 we report the test for the null hypothesis of constant returns to scale (CRS), i.e. $H_0 : \beta_1 + \beta_2 = 1$, as well as the test for the homogeneity assumption, both on the basis of the $J$-test for the validity of over-identifying restrictions and the $F$-test of equality of coefficients in the longitudinal and singleton samples in the fully interacted OLS regression. The homogeneity assumption is not rejected.

The first visible result is that the standard error of the GMM estimate is always smaller than that of the FE estimate. The gain of efficiency is approximately 8-9 percent, a non-negligible improvement for this type of TFP analysis. In one instance, this increases the level of significance of the coefficient estimation (for the elasticity to capital, in the model with time dummy). Estimates of the elasticities to labor and capital are in line with the findings of the existing literature.

## 5. Conclusions

We devise an innovative procedure, built within a GMM framework, that exploits unpaired observations (singletons) to increase the efficiency of FE estimates. Longitudinal data allow tackling bias due to the correlation between variables of interest and unobserved time-invariant characteristics. The use of the singletons allows reducing the standard errors of the FE estimates, potentially increasing their significance.

Our procedure relies on the plausible assumption that the relationship between observed and unobserved characteristics is homogeneous across longitudinal and singleton samples. The assumption can be easily tested.

We find efficiency gains through a set of Monte Carlo experiments. We then apply our procedure to the estimation of a TFP regression. Estimation efficiency increases in all model specifications, in the order of 8-9 percent.

---

quartile range ($iqr$), and considered as outliers those observations with values outside the interval defined by $m \pm 1.5iqr$. We removed about 20% of observations.

[10] The [8] correction has been applied for the computation of the standard errors.

| Variable | FE | GMM | FE | GMM |
|---|---|---|---|---|
| $\ln L$ | .653*** | .650*** | .607*** | .602*** |
| | (.121) | (.116) | (.131) | (.126) |
| $\ln K$ | .156*** | .166*** | .126** | .132*** |
| | (.057) | (.050) | (.053) | (.048) |
| Time dummy | no | no | yes | yes |
| Test CRS | 2.37 | 2.16 | 3.77 | 3.80 |
| [p-value] | [.124] | [.142] | [.052] | [.051] |
| $J$-test | | 1.34 | | 1.49 |
| [p-value] | | [.719] | | [.828] |
| $F$-test | | .462 | | .381 |
| [p-value] | | [.709] | | [.823] |

Statistical significance: * 10%, ** 5%, *** 1%.

Table 2: Results of the econometric estimation; standard errors (in parenthesis) are computed by applying the [8] correction; p-value of reported test statistics among squared brackets

## References


[1] Amemiya, T., and MaCurdy, T.E. (1986): "Instrumental-Variable Estimation of an Error-Components Model", *Econometrica* **54**(4), 869-880.

[2] Breusch, T.S., Mizon, G.E., and Schmidt, P. (1989): "Efficient Estimation Using Panel Data", *Econometrica* **57**(3), 695-700.

[3] 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.

[4] Commander, S., and Svejnar, J. (2011): "Business Environment, Exports, Ownership, and Firm Performance", *The Review of Economics and Statistics*, **93**(1), 309-337.

[5] Hansen, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica* **50**(4), 1029-1054. doi: 10.2307/1912775

[6] Hausman, J.A., and Taylor, W.E. (1981): "Panel Data and Unobservable Individual Effects", *Econometrica* **49**(6), 1377-1398.

[7] Verbeek, M. (2008): *A Guide to Modern Econometrics*, John Wiley & Sons.

[8] Windmeijer, F. (2005): A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators, *Journal of econometrics* **126**(1), 25-51.