Title: Use information from singletons in fixed effect estimation: *xtfesing*

Authors:

- Laura Magazzini, Department of Economics, University of Verona, Italy; e-mail: laura.magazzini@univr.it, corresponding author.
- Randolph Luca Bruno, School of Slavonic and East European Studies, University College London, UK, Fondazione Rodolfo DeBenedetti and Institute for the Study of Labor (IZA)
- Marco Stampini, Social Protection and Health Division, Inter-American Development Bank, USA

Abstract. This article describes the *xtfesing* command. The command implements a GMM estimator that allows exploiting singleton information in fixed effect panel data regression as in Bruno, Magazzini and Stampini (2020).

Keywords. Panel data, fixed effects, singletons, estimation efficiency.

1. Introduction

Analysis of longitudinal (panel) data has the advantage of allowing consistent estimation of the model parameters even in the presence of unobserved heterogeneity, i.e. decreasing the risk of omitted variables bias. The fixed effect approach (in STATA *xtreg* command with the *fe* option) allows estimating the effect of time-varying variables even in the presence of correlation with the error term, provided that the correlation is driven by omitted time-invariant variables, either observed or unobservable (such as individual preferences or gender, firms’ propensity to patent or foundation year, etc.). Consistent estimation of the parameters of interest is obtained by using the within-group transformation that removes the individual average from the variables included in the model. Singleton units, i.e. those units observed only at one point in time, do not contribute to the analysis, as their within-group transformation is identically equal to zero.

While most textbook examples consider a balanced panel data set, real data often entail an unbalanced set of units, with a substantial share of singleton observations. In some cases, singletons are due to natural enterprise mortality and refreshment of the sample with new units. This type of attrition is common in databases like Orbis (https://www.bvdinfo.com/en-gb) or the Business Environment and Enterprise Performance Survey (https://www.ebrd.com/data; https://www.enterprisesurveys.org/). In the case of rotating panels, singletons are the result of the sampling framework. This happens in many labor force surveys in which a share of the observations is replaced in each wave, and the observations that are interviewed only in the first wave are singletons by design. Attrition and singletons can also be due to the death of part of the sample. This is particularly relevant for samples of older people, as in the United States’ Health and Retirement Study (https://hrs.isr.umich.edu/about) or the Mexican Health and Aging Study (http://www.mhasweb.org/). Migration and non-response are other common causes of attrition and the resulting presence of singleton observations in longitudinal data.

In this paper we describe the *xtfesing* command, that estimates a static panel data model with fixed effects and exploits information from the singleton units in the sample with the aim to increase estimation efficiency. The methodology has been proposed by Bruno, Magazzini and Stampini (2020;
henceforth BMS20). The method can also be used to “pool” panel datasets and cross-section observations from other survey waves as in Bruno and Stampini (2009).

xtfesing implements a two-step GMM estimator (Hansen, 1982). Its validity relies on the homogeneity assumption: it requires that the OLS bias is the same for the panel units and the singletons.

The paper proceeds as follow. Section 2 describes the methodology. Section 3 presents the syntax of the xtfesing command, its estimation options, and its post estimation characteristics. An example based on the STATA dataset “nlswork” is provided in Section 4.

2. Method
Consider the linear static panel data model with individual effects \((i = 1, \ldots, N; t = 1, \ldots, T)\):

\[
y_{it} = x_{it}' \beta + u_i + e_{it}
\]

(1)

where \(y_{it}\) represents the dependent variable of interest measured on unit \(i\) at time \(t\), \(x_{it}\) a \(k \times 1\) vector of observable characteristics of unit \(i\) at time \(t\) (an intercept can be included), \(\beta\) a \(k \times 1\) vector of parameters to be estimated, \(u_i\) the individual effect and \(e_{it}\) the idiosyncratic component. The variables in \(x_{it}\) are allowed to be arbitrarily correlated with \(u_i\), but the assumption of strict exogeneity is imposed so that correlation of \(x_{it}\) with \(e_{it}\) is ruled out at any time \((s = 1, \ldots, T)\). The panel can be unbalanced: the number of time period observations for unit \(i\) equals \(T_i\).

In the set-up of Model (1), the fixed effect estimator is consistent: the presence of an unbalanced\(^1\) panel only complicates the notation, but does not affect the properties of the estimator.

Define \(\bar{x}_{j, it} = x_{j, it} - \bar{x}_{j, it}\) with \(\bar{x}_{j, it} = \sum_t x_{j, it} / T_i\) \((j = 1, \ldots, k)\), the individual demeaned independent variables. In the case of \(T_i = 1\) (singleton units), \(\bar{x}_{j, it} = 0\) for each regressor \(j\). The fixed effect estimator can be obtained as an instrumental variable estimator of Model (1) with instruments \(\bar{x}_{j, it}\). The following \(k\) moment conditions are therefore satisfied (see eq. 2 in BMS20):\(^2\)

\[
E[\bar{x}_{it}(y_{it} - x_{it}' \beta)] = 0
\]

(2)

In contrast, due to the possibility of correlation between the independent variables and the individual component \(u_i\), the OLS estimator may be biased. Denote with \(b\) the OLS bias, also the following moment conditions are satisfied (see eq. 3 in BMS20):

\[
E[x_{it}(y_{it} - x_{it}' (\beta + b))] = 0
\]

(3)

As an equal number of moment conditions and parameters is added, the estimated coefficients in \(\beta\) are unaffected. However, information from singleton units can be further exploited in order to obtain efficiency gains under the assumption that the OLS bias is the same for the singletons and those units that are observed more than once. Denote with \(i = s\) the singletons: the following moment condition can also be considered (see eq. 4 in BMS20):

\[
E[x_{st}(y_{st} - x_{st}' (\beta + b))] = 0
\]

(4)

\(^1\) The nature of “unbalance” should be random and not systematic, though.\(^2\) If an intercept is included in the model, the corresponding variable in \(x_{it}\) should not be demeaned.
We propose a GMM estimator based on moment conditions (2), (3), and (4). The computation considers a two-step procedure based on the \texttt{gmm} STATA command with clustered standard errors (cluster defined on the basis of the group variable that identifies the units). It includes the Windmeijer (2005)’s formula for the correction of the two-step estimated standard error.

The assumption of homogeneity can be tested using a regression framework or on the basis of the test of over-identifying conditions based on the value of the minimized GMM criterion. The two test statistics are provided with the proposed command. Please refer to BMS20 for details.

3. The \texttt{xtfesing} command

The syntax of the \texttt{xtfesing} command is as follows

\begin{verbatim}
xtfesing depvar [indepvars] [if] [in] [, id(varname) nowindmeijer level(#)]
\end{verbatim}

where \texttt{depvar} represents the dependent variable and \texttt{indepvars} the list of independent variables. A subsample of the data can be specified using the \texttt{if} condition or in range, as usual.

The following options are available:

- \texttt{id(varname)} with the variable \texttt{varname} identifying the grouping variable. The option can be omitted when the variables identifying the panel dimensions have been specified with the \texttt{xtset} command. In this case the variable identifying the panel units is considered (if the option is omitted but no \texttt{xtset} command has been defined before \texttt{xtfesing}, an error message is displayed);
- \texttt{nowindmeijer}: by default, the standard error produced by \texttt{xtfesing} are computed using the Windmeijer (1995)’s correction. When the \texttt{nowindmeijer} option is specified, the default standard errors computed by the STATA’s \texttt{gmm} command are reported;
- \texttt{level(#)} specifies the confidence level. The default value is 95 (95%).

The \texttt{xtfesing} command allows the use of the post-estimation command \texttt{predict}. The following options can be specified:

- \texttt{xb} \quad a + xb, fitted values (the default)
- \texttt{ue} \quad u_i + e_it, the combined residual

The \texttt{xtfesing} command stores the following results in e():

- Scalars:
  - e(rank) \quad rank of e(V)
  - e(N) \quad number of observations
  - e(Q) \quad value of minimized GMM criterion
  - e(J) \quad value of J-test of overidentifying restrictions
  - e(J_df) \quad degrees of freedom of J-test
e(converged) 1 if converged, 0 otherwise
e(N_eq) number of equations passed to gmm command, equal to 3
e(k) number of estimated parameters
e(n_moments) number of moment conditions
e(N_clust) number of clusters
e(F_hom) value of F statistic for regression-based test of homogeneity
e(F_hom_p) p-value of F statistics for homogeneity
e(NS) number of singletons

- Macros:
e(cmd) xtfesing
e(cmdline) command line, as typed by the user
e(depvar) name of the dependent variable
e(rhs) list of the independent variable(s)
e(predict) xtfesing_p, name of the command used for predict
e(clustvar) name of clustering variable, also used to identify singletons
e(vcetype) Robust
e(vce) cluster
e(wmatrix) name of clustvar, equal to varname in the id() option
e(estimator) twostep
e(winit) Unadjusted
e(nocommonesample) nocommonesample
e(properties) b V

- Matrices:
e(b) vector of the estimated coefficients
e(V) variance-covariance matrix of the coefficients
e(Vunc) uncorrected variance-covariance matrix of the coefficients,
if e(V) computed according to Windmeijer (1995)
e(W) weight matrix used for final round of estimation
e(S) moment covariance matrix used in robust VCE computations
e(init) initial values of the estimator
4. Example: a wage equation

We consider the dataset \textit{nlswork}, available online from the STATA website:\textsuperscript{3}

\texttt{. webuse nlswork}

The dataset contains information on young women who were between the age of 14 and 26 in 1968. Data are extracted from the National Longitudinal Surveys (NLS) conducted by the U.S. Department of Labor.

We specify the panel dimensions by using the \texttt{xtset} command:

\texttt{. xtset idcode year}
\begin{verbatim}
    panel variable:  idcode (unbalanced)
    time variable:  year, 68 to 88, but with gaps
    delta:  1 unit
\end{verbatim}

The dataset contains 4711 units observed over 15 time periods (from 1968 to 1988, with some gaps). The panel is unbalanced: a description of the dataset structure with \texttt{xtdescribe} yields the following results:

\texttt{. xtdescribe}

\begin{verbatim}
idcode:  1, 2, ..., 5159                      n =  4711
year:  68, 69, ..., 88                        T =  15
Delta(year) = 1 unit
Span(year)  = 21 periods
(idcode*year uniquely identifies each observation)
\end{verbatim}

Distribution of $T_i$: 
\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Freq. & Percent & Cum. \% & Pattern \\
\hline
136 & 2.89 & 2.89 & 1.................. \\
114 & 2.42 & 5.31 & .................1 \\
89 & 1.89 & 7.20 & ..................1.11 \\
87 & 1.85 & 9.04 & ..................11 \\
86 & 1.83 & 10.87 & 111111.1.11.1.11.1.11 \\
61 & 1.29 & 12.16 & ..................11.11 \\
56 & 1.19 & 13.35 & 11.................. \\
54 & 1.15 & 14.50 & ..................1.11 \\
54 & 1.15 & 15.64 & ........1.11.1.11.1.11 \\
3974 & 84.36 & 100.00 & (other patterns) \\
\hline
4711 & 100.00 & | XXXXX.X.XX.X.XX.X.XX
\hline
\end{verbatim}

The two most common patterns are indeed singletons: 136 units are observed only in the first time period, and 114 are observed only in the last time period. Singletons also include units with a single

\textsuperscript{3} We are running the example on STATA 15 so that the dataset is drawn from www.stata-press.com/data/r15.
observation at any intermediate time, plus units with more than one observation that enter the estimation sample only once due to missing values in the variables considered by the model. This last group is not counted with \texttt{xtdescribe} which is based on the number of lines occupied by each unit in the data set.

We consider the logarithm of wage (\texttt{ln Wage}) as dependent variable and include among the independent variables total work experience (\texttt{ttl exp}) and its square, a dummy variable for union membership (\texttt{union}), the age of the woman, and three dummy variables to identify her residence (\texttt{south}, \texttt{c city}, and \texttt{not_smsa}).

We first generate the square of the variable \texttt{ttl exp}:

\begin{verbatim}
. gen ttl_exp2 = ttl_exp^2
\end{verbatim}

As a benchmark for the proposed estimation procedure, we also consider the fixed effect estimator. Robust standard error, clustered over \texttt{idcode} are considered to account for the possibility of heteroskedasticity and autocorrelation in the idiosyncratic component. Some missing values are present so that the number of units decreases to 4150.\footnote{Validity of panel data estimators with unbalanced datasets relies on the assumption that observability is not due to endogenous reasons. In particular, the fixed effect estimator would not be affected by selectivity bias if selection is dependent upon the individual effect \(u_i\). In this framework, selection can also depend on the idiosyncratic component \(e_u\), provided that the relationship is time invariant (Verbeek, 2004, p. 383).}

\begin{verbatim}
. xtreg ln_wage ttl_exp* union age south c_city not_smsa , fe cluster(idcode)
\end{verbatim}

\begin{tabular}{l r r r r r}
 & Coef. & Std. Err. & t & P>|t| & \([95\% \text{ Conf. Interval}]\) \\
----------------- & ------ & ---------- & --- & ----- & ------------------- \\
\texttt{ln wages} & 0.0653815 & 0.0038493 & 16.99 & 0.000 & 0.0578348 - 0.0729282 \\
\texttt{ttl exp} & -0.000965 & 0.000127 & -7.60 & 0.000 & -0.001214 - -0.0007161 \\
\texttt{union} & 0.0961601 & 0.0093992 & 10.23 & 0.000 & 0.0777326 - 0.1145876 \\
\texttt{age} & -0.0180308 & 0.0018058 & -9.99 & 0.000 & -0.0215711 - -0.144905 \\
\texttt{south} & -0.0649143 & 0.0212538 & -3.05 & 0.002 & -0.1065831 - -0.0232455 \\
\texttt{c city} & 0.0067234 & 0.0122647 & 0.55 & 0.584 & 0.001722 - 0.0307689 \\
\texttt{not_smsa} & -0.0888541 & 0.0190039 & -4.68 & 0.000 & -0.1261118 - -0.0515964 \\
\texttt{cons} & 1.920127 & 0.0401127 & 47.87 & 0.000 & 1.841485 - 1.99877 \\
\end{tabular}

\begin{verbatim}
sigma_u | 0.36937539
sigma_e | 0.25428694
\rho | 0.67845928 (fraction of variance due to \texttt{u}_i)
\end{verbatim}
Overall, the estimation sample includes 665 singletons: the presence of singletons is reflected in the number of years of observations, which ranges from 1 to 12.

The same equation is estimated using the BMS20 procedure implemented with the *xtfesing* command:

```
.xtfesing ln_wage ttl_exp* union age south c_city not_smsa
```

GMM estimation results

<table>
<thead>
<tr>
<th>Total number of observations</th>
<th>19226</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of units</td>
<td>4150</td>
</tr>
<tr>
<td>Number of singletons</td>
<td>665</td>
</tr>
<tr>
<td>(6.02% of total n. of units)</td>
<td></td>
</tr>
</tbody>
</table>

(Std. Err. adjusted for 4,150 clusters in idcode)

| ln_wage | Coef. | Std. Err. | z   | P>|z| | [95% Conf. Interval] |
|---------|-------|-----------|-----|-----|----------------------|
| beta    |       |           |     |     |                      |
| ttl_exp | .0661623 | .0038393 | 17.23 | 0.000 | .0586374 | .0736873 |
| ttl_exp2| -.0009941 | .001264  | -7.86 | 0.000 | -.0012419 | -.0007464 |
| union   | .0969912  | .0093628 | 10.36 | 0.000 | .0786405  | .1145324 |
| age     | -.0179975 | .0017986 | -10.01| 0.000 | -.0215226 | -.144724  |
| south   | -.0622753 | .0212204 | -2.94 | 0.003 | -.1038469 | -.0207036 |
| c_city  | .0079747  | .0122257 | 0.65  | 0.514 |                      |          |
| not_smsa| -.0885119 | .0189696 | -4.67 | 0.000 | -.1256915 | -.0513322 |
| _cons   | 1.913807  | .0401157 | 47.71 | 0.000 | 1.835183  | 1.992432  |

| bias    |       |           |     |     |                      |
|---------|-------|-----------|-----|-----|                      |
| ttl_exp | .0040013 | .0041352 | 0.97 | 0.333 | -.0041036 | .0121062  |
| ttl_exp2| -.0002135 | .0001517 | -1.41 | 0.159 | -.0005108 | .000838   |
| union   | .0600835 | .012069  | 4.98 | 0.000 | .0364364 | .0837305  |
| age     | .0064886 | .0018699 | 3.47 | 0.001 | .0028239 | .0101532  |
| south   | -.0755591 | .0225083 | -3.36 | 0.001 | -.1197065 | -.0314756 |
| c_city  | -.0333657 | .0150273 | -2.22 | 0.026 | -.0628186 | -.0039127 |
| not_smsa| -.1280753 | .0212832 | -6.02 | 0.000 | -.1697896 | -.086361  |
| _cons   | -.1523933 | .0412182 | -3.70 | 0.000 | -.2331795 | -.0716072 |

The option *id()* is omitted because we previously defined the panel through the command *xtset*. The variable *idcode* is therefore considered to identify the units.

At the top of the table of results, we have information on the total number of observations (19226), the total number of units (4150) and the number of singletons (665, corresponding to 16.02% of the total number of units).

The table of results reports the estimated coefficients for “beta” (the consistent estimator of the coefficient of interest) and the OLS “bias” for each variable in the estimated equation. Note that when
the `predict` command is invoked after `xtfesing`, only the coefficients in “beta” are considered for computing predicted values and residuals (coefficients in “bias” are not included in the computations).

At the bottom, the table reports the two tests of the homogeneity assumption, required for the validity of the proposed approach:

- The Hansen-based test of homogeneity, corresponding to the test of overidentifying restrictions for the GMM estimation, produces a value of 12.68 with a p-value of 0.123;
- The regression-based test of homogeneity produces a value of 1.69 with a p-value of 0.096.

Both tests do not reject the null hypothesis of homogeneity at the 5% level of significance, so that the BMS20 procedure can be applied to these data.

In this specific case, the reduction in the standard errors is limited (or null). As pointed in BMS20, efficiency gains can be negligible with a long time dimension or when the share of singleton is not substantial.

For illustration purposes, we limit the analysis to the last three years of the dataset (85, 87, and 88). We also restrict the sample, and only include white women. In this way, we “artificially” generate a dataset characterized by a small time dimension and a larger (even though, still fairly limited) share of singletons.

```
xtreg ln_wage ttl_exp* union age south c_city not_smsa if year>=85 & race==1, fe cluster(idcode)
Fixed-effects (within) regression                          Number of obs     =      4,408
Group variable: idcode                                         Number of groups  =      2,053
R-sq:                                               Obs per group:
       within = 0.0749                                          min =          1
       between = 0.2816                                       avg =        2.1
       overall = 0.2561                                        max =          3
corr(u_i, Xb) = 0.0353                                      F(7,2052)         =      24.13
                                                                     Prob > F          =     0.0000
(Std. Err. adjusted for 2,053 clusters in idcode)

| ln_wage | Coef.    | Robust Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|----------|------------------|-------|------|---------------------|
| ttl_exp | 0.0856074 | 0.0158313        | 5.41  | 0.000 | 0.0545604 1.1166544 |
| ttl_exp2 | -0.0014964 | 0.0003506      | -4.27 | 0.000 | -0.0021841 -0.0008088 |
| union   | 0.0837033  | 0.0210204       | 3.98  | 0.000 | 0.0424798 0.1249267 |
| age     | -0.0142388 | 0.0115589       | -1.23 | 0.218 | -0.0369072 0.0084295 |
| south   | -0.0560606 | 0.0671243       | -0.84 | 0.404 | -0.1876994 0.0755782 |
| c_city  | 0.0454149  | 0.0353415       | 1.29  | 0.199 | -0.023894 0.1147238 |
| not_smsa | -0.0777794 | 0.0458192      | -1.70 | 0.090 | -0.1676364 0.0120776 |
| _cons   | 1.685030   | 0.3042241       | 5.54  | 0.000 | 1.08841 2.28165 |
```

```
sigma_u | 0.4272089
sigma_e | 0.20786549
rho     | 0.80857291 (fraction of variance due to u_i)
```
GMM estimation results

Total number of observations 4408
Total number of units 2053
Number of singletons 573 (27.91% of total n. of units)

(Std. Err. adjusted for 2,053 clusters in idcode)

|                        | Coef.  | Robust Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|------------------------|--------|------------------|-------|-------|---------------------|
| ln_wage                |        |                  |       |       |                     |
| ttl_exp                | .0864941 | .0157324         | 5.50  | 0.000 | .0556592 -.1173289 |
| ttl_exp2               | -.0014791 | .0003499        | -4.23 | 0.000 | -.0021649 -.0007933 |
| union                  | .0850271 | .0209337         | 4.06  | 0.000 | .0439977 .1260565  |
| age                    | -.0157543 | .0115209         | -1.37 | 0.171 | -.0383348 .0068263 |
| south                  | -.0565427 | .0669068         | -0.85 | 0.398 | -.1876775 .0745922 |
| c_city                 | .0440417 | .0352062         | 1.25  | 0.211 | -.0249611 .1130446 |
| not_smsa               | -.0814644 | .0457795        | -1.78 | 0.075 | -.1711906 .0082619 |
| _cons                  | 1.727003 | .3030677         | 5.70  | 0.000 | 1.133001 2.321005  |

bias

|                        | Coef.  | Robust Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|------------------------|--------|------------------|-------|-------|---------------------|
| ttl_exp                | .0010469 | .0173146         | 0.06  | 0.952 | -.0328892 .034983  |
| ttl_exp2               | -.0001146 | .0004621        | -0.25 | 0.804 | -.0010203 .0007912 |
| union                  | .0664312 | .0277637         | 2.39  | 0.017 | .0120153 1.120847  |
| age                    | .0076781 | .0116198         | 0.66  | 0.509 | -.0150962 .0304525 |
| south                  | .0309872 | .068533          | 0.45  | 0.651 | -.103335 1.1653093 |
| c_city                 | -.0289911 | .041279         | -0.70 | 0.482 | -.1098965 .0519142 |
| not_smsa               | -.137757 | .0481799         | -2.86 | 0.004 | -.2321879 -.0433261 |
| _cons                  | -.2587639 | .3101621        | -0.83 | 0.404 | -.8666705 .3491426 |

Hansen-based test of homogeneity: J = 16.86 (p-value = 0.032)
Regression-based test of homogeneity: F = 2.21 (p-value = 0.024)

In this case, standard errors tend to be lower when using xtfesing as compared to xtregr. The homogeneity assumption is not rejected at the 1% level of significance.

BMS20 considers cases in which the share of singletons reaches or exceeds 50%. They show that, in those cases, the procedure implemented by xtfesing leads to large improvements in estimation efficiency.

References


**About the authors**

Laura Magazzini is Associate Professor of Econometrics at the Department of Economics, University of Verona, Italy. She obtained the PhD in 2004 at the Sant’Anna of Advanced Studies in Pisa, Italy. She graduated in Economics and Statistics at the University of Florence. During her studies, she was visiting scholar at Carnegie Mellon University, Pittsburgh, USA and the University of Sydney, Sydney, Australia. Her research interests are centered around microeconometrics, industrial economics, the economics of innovation, competition policy and econometric methods, with particular reference to panel data analysis.

Randolph Luca Bruno is Associate Professor of Economics at University College London, SSEES. He holds visiting positions at the London School of Economics and Political Science, Università della Svizzera Italiana, University of Bari and affiliations at IZA-Bonn –Research Fellow- and Fondazione Rodolfo DeBenedetti-Milan –Senior Research Fellow-. His main research interests revolve around applied micro-econometrics, institutional/comparative economics, labor economics and innovation both from a Macro as well as Micro perspective.

Marco Stampini is Social Protection Lead Specialist at the Interamerican Development Bank in Washington DC. He focuses on the design, implementation and evaluation of social protection policies and programs in Latin America and the Caribbean. In recent times, his studies have focused on conditional cash transfers and long-term care. He holds a Masters in Applied Economics from CORIPE Piedmont, and a Ph.D. in Environmental Economics from Sant’Anna School of Advanced Studies.