The Stata Journal (2020)
20, Number 4, pp. 965–975 DOI: 10.1177/1536867X20976326

Using information from singletons in fixed-effects estimation: xtfesing

Laura Magazzini
Institute of Economics
Sant'Anna School of Advanced Studies
Pisa, Italy
laura.magazzini@santannapisa.it

Randolph Luca Bruno
School of Slavonic and East European Studies
University College London
London, UK
Fondazione Rodolfo DeBenedetti and Institute for the Study of Labor (IZA)

Marco Stampini
Social Protection and Health Division
Inter-American Development Bank
Washington, DC

Abstract. In this article, we describe the `xtfesing` command. The command implements a generalized method of moments estimator that allows exploiting singleton information in fixed-effects panel-data regression as in Bruno, Magazzini, and Stampini (2020, *Economics Letters* 186: Article 108519).

Keywords: st0623, xtfesing, 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, that is, of decreasing the risk of omitted variables bias. The fixed-effects approach (in Stata, the `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 or firms' propensity to patent or foundation year). 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, that is, those units observed only at one point in time, do not contribute to the analysis, because their within-group transformation is identically equal to zero.

While most textbook examples consider a balanced panel dataset, 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.beeps-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 nonresponse are other common causes of attrition and the resulting presence of singleton observations in longitudinal data.

In this article, we describe the `xtfesing` command, which 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). 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 generalized method of moments (GMM) estimator (Hansen 1982). Its validity relies on the homogeneity assumption: it requires that the ordinary least-squares (OLS) bias be the same for the panel units and the singletons.

The article proceeds as follow. Section 2 describes the methodology. Section 3 presents the syntax of the `xtfesing` command, its estimation options, and its postestimation characteristics. Section 4 provides an example based on the Stata dataset `nlswork.dta`.

2 Method

Consider the linear static panel-data model with individual effects ($i = 1, \ldots, N$; $t = 1, \ldots, T_i$),

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + e_{it} \tag{1}$$

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

In the setup of (1), the fixed-effects estimator is consistent: the presence of an unbalanced[1] panel complicates only the notation but does not affect the properties of the estimator.

Define $\ddot{x}_{j,it} = x_{j,it} - \overline{x}_{j,i}$ with $\overline{x}_{j,i} = \Sigma_t x_{j,it}/T_i$ ($j = 1,\dots,k$), the individual demeaned independent variables. In the case of $T_i = 1$ (singleton units), $\ddot{x}_{j,it} = 0$ for each regressor j. The fixed-effects estimator can be obtained as an instrumental variable estimator of (1) with instruments $\ddot{x}_{j,it}$. The following k moment conditions are therefore satisfied [see (2) in Bruno, Magazzini, and Stampini (2020)]:[2]

$$E\left\{\ddot{\mathbf{x}}_{it}\left(y_{it} - \mathbf{x}'_{it}\boldsymbol{\beta}\right)\right\} = 0 \tag{2}$$

In contrast, because of the possibility of correlation between the independent variables and the individual component u_i, the OLS estimator may be biased. Denote with $\mathbf{b}$ the OLS bias; also, the following moment conditions are satisfied [see (2) in Bruno, Magazzini, and Stampini (2020)]:

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

As an equal number of moment conditions and parameters are added, the estimated coefficients in $\boldsymbol{\beta}$ are unaffected. However, information from singleton units can be further exploited 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 (3) in Bruno, Magazzini, and Stampini (2020)]:

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

We propose a GMM estimator based on moment conditions (2), (2), and (3). The computation considers a two-step procedure based on the `gmm` Stata command with clustered standard errors (cluster defined on the basis of the group variable that identifies the units). It includes Windmeijer's (2005) 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 overidentifying conditions based on the value of the minimized GMM criterion. The two test statistics are provided with the proposed command. Please refer to Bruno, Magazzini, and Stampini (2020) for details.

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 $\mathbf{x}_{it}$ should not be demeaned.

3 The xtfesing command

3.1 Syntax

The syntax of the `xtfesing` command is as follows:

`xtfesing` *depvar* [*indepvars*] [*if*] [*in*] [`,` `id(`*varname*`)` `nowindmeijer` `level(`*#*`)`]

depvar represents the dependent variable, and *indepvars* the list of independent variables. A subsample of the data can be specified using the `if` or `in` qualifier, as usual.

3.2 Options

`id(`*varname*`)` specifies *varname* identifying the grouping variable. The option can be omitted when the variables identifying the panel dimensions have been specified with the `xtset` command. In this case, the variable identifying the panel units is considered (if the option is omitted but no `xtset` command has been defined before `xtfesing`, an error message is displayed).

`nowindmeijer` specifies that the default standard errors computed by Stata's `gmm` command be reported. By default, they are computed using Windmeijer's (2005) correction.

`level(`*#*`)` specifies the confidence level. The default is `level(95)`.

3.3 Postestimation command

The `xtfesing` command allows the use of the postestimation command `predict`. The following options can be specified:

`xb`	$a + \mathbf{x}'\boldsymbol{\beta}$, fitted values (the default)
`ue`	$u_i + e_{it}$, the combined residual

3.4 Stored results

`xtfesing` stores the following results in `e()`:

Scalars	
`e(N)`	number of observations
`e(k)`	number of estimated parameters
`e(N_clust)`	number of clusters
`e(rank)`	rank of `e(V)`
`e(converged)`	`1` if converged, `0` otherwise
`e(Q)`	value of minimized GMM criterion
`e(J)`	value of J-test of overidentifying restrictions
`e(J_df)`	degrees of freedom of J-test
`e(N_eq)`	number of equations passed to `gmm` command, equal to three
`e(n_moments)`	number of moment conditions
`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 as typed
`e(depvar)`	name of the dependent variable
`e(rhs)`	list of the independent variables
`e(clustvar)`	name of clustering variable; also used to identify singletons
`e(vce)`	`cluster`
`e(vcetype)`	`Robust`
`e(predict)`	`xtfesing_p`, name of the command used for `predict`
`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 (2005)
`e(W)`	weight matrix used for final round of estimation
`e(S)`	moment covariance matrix used in robust variance–covariance estimators computations
`e(init)`	initial values of the estimator

4 Example: A wage equation

We consider `nlswork.dta`, available online from the Stata website:[3]

```
. webuse nlswork
(National Longitudinal Survey.  Young Women 14-26 years of age in 1968)
```

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

3. We are running the example on Stata 16, so the dataset is drawn from https://www.stata-press.com/data/r16.

We specify the panel dimensions by using the `xtset` command:

```
. xtset idcode year
       panel variable:  idcode (unbalanced)
        time variable:  year, 68 to 88, but with gaps
                delta:  1 unit
```

The dataset contains 4,711 units observed over 15 time periods (from 1968 to 1988, with some gaps). The panel is unbalanced: a description of the dataset structure with `xtdescribe` yields the following results:

```
. xtdescribe
  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)
Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         5         9      13      15
     Freq.  Percent    Cum. |  Pattern
 ---------------------------+----------------------
      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 |  ..............11.1.11
       56      1.19   13.35 |  11...................
       54      1.15   14.50 |  ...............1.1.11
       54      1.15   15.64 |  .......1.11.1.11.1.11
     3974     84.36  100.00 | (other patterns)
 ---------------------------+----------------------
     4711    100.00         |  XXXXXX.X.XX.X.XX.X.XX
```

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 observation at any intermediate time, plus units with more than one observation that enter the estimation sample only once because of missing values in the variables considered by the model. This last group is not counted with `xtdescribe`, which is based on the number of lines occupied by each unit in the dataset.

We consider the logarithm of wage (`ln_wage`) as dependent variable and include among the independent variables total work experience (`ttl_exp`) and its square, a dummy variable for union membership (`union`), the age of the woman, and three dummy variables to identify her residence (`south`, `c_city`, and `not_smsa`).

We first generate the square of the variable `ttl_exp`:

```
. generate ttl_exp2 = ttl_exp^2
```

As a benchmark for the proposed estimation procedure, we also consider the fixed-effects estimator. Robust standard error, clustered over `idcode`, is considered to account for the possibility of heteroskedasticity and autocorrelation in the idiosyncratic component. Some missing values are present, so the number of units decreases to 4,150.[4]

```
. xtreg ln_wage ttl_exp* union age south c_city not_smsa, fe cluster(idcode)
Fixed-effects (within) regression               Number of obs     =     19,226
Group variable: idcode                          Number of groups  =      4,150

R-sq:                                           Obs per group:
     within  = 0.1501                                         min =          1
     between = 0.2892                                         avg =        4.6
     overall = 0.2364                                         max =         12

                                                F(7,4149)         =     179.70
corr(u_i, Xb)  = 0.1227                         Prob > F          =     0.0000

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

                             Robust
     ln_wage       Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]

     ttl_exp    .0653815   .0038493    16.99   0.000     .0578348    .0729282
    ttl_exp2    -.000965    .000127    -7.60   0.000    -.001214   -.0007161
       union    .0961601   .0093992    10.23   0.000     .0777326    .1145876
         age   -.0180308   .0018058    -9.99   0.000    -.0215711   -.0144905
       south   -.0649143   .0212538    -3.05   0.002    -.1065831   -.0232455
      c_city    .0067234   .0122647     0.55   0.584     -.017322    .0307689
    not_smsa   -.0888541   .0190039    -4.68   0.000    -.1261118   -.0515964
       _cons    1.920127   .0401127    47.87   0.000     1.841485     1.99877

     sigma_u   .36937539
     sigma_e   .25428694
         rho   .67845928   (fraction of variance due to u_i)
```

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.

4. Validity of panel data-estimators with unbalanced datasets relies on the assumption that observability is not due to endogenous reasons. In particular, the fixed-effects 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_{it}, provided that the relationship is time invariant (Verbeek 2004, 383).

The same equation is estimated using the Bruno, Magazzini, and Stampini (2020) procedure implemented with the `xtfesing` command:

```
. xtfesing ln_wage ttl_exp* union age south c_city not_smsa
GMM estimation results
Total number of observations     19226
       Total number of units      4150
        Number of singletons       665 (16.02% of total n. of units)
                           (Std. Err. adjusted for 4,150 clusters in idcode)

                             Robust
     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   .0001264    -7.86   0.000    -.0012419   -.0007464
       union    .0969912   .0093628    10.36   0.000     .0786405     .115342
         age   -.0179975   .0017986   -10.01   0.000    -.0215226   -.0144724
       south   -.0622753   .0212104    -2.94   0.003    -.1038469   -.0207036
      c_city    .0079747   .0122257     0.65   0.514    -.0159872    .0319366
    not_smsa   -.0885119   .0189696    -4.67   0.000    -.1256915   -.0513322
       _cons    1.913807   .0401152    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    .0000838
       union    .0600835    .012065     4.98   0.000     .0364364    .0837305
         age    .0064886   .0018698     3.47   0.001     .0028239    .0101532
       south    -.075591   .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

Hansen-based test of homogeneity:       J =     12.68 (p-value =     0.123)
Regression-based test of homogeneity:   F =      1.69 (p-value =     0.096)
```

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 (19,226), the total number of units (4,150) 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 the Bruno, Magazzini, and Stampini (2020) procedure can be applied to these data.

In this case, the reduction in the standard errors is limited (or null). As Bruno, Magazzini, and Stampini (2020) point out, efficiency gains can be negligible with a long time dimension or when the share of singletons is not substantial.

For illustration, we limit the analysis to the last three years of the dataset (85, 87, and 88). We also restrict the sample by only including 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

                                                F(7,2052)         =      24.13
corr(u_i, Xb)  = 0.0353                         Prob > F          =     0.0000

                             (Std. Err. adjusted for 2,053 clusters in idcode)
------------------------------------------------------------------------------
             |               Robust
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ttl_exp |   .0856074   .0158313     5.41   0.000     .0545604    .1166544
    ttl_exp2 |  -.0014964   .0003506    -4.27   0.000    -.0021841   -.0008088
       union |   .0837033   .0210204     3.98   0.000     .0424798    .1249267
         age |  -.0142388   .0115589    -1.23   0.218    -.0369072    .0084295
       south |  -.0560606   .0671243    -0.84   0.404    -.1876994    .0755782
      c_city |   .0454149   .0353415     1.29   0.199     -.023894    .1147238
    not_smsa |  -.0777794   .0458192    -1.70   0.090    -.1676364    .0120776
       _cons |    1.68503   .3042241     5.54   0.000      1.08841     2.28165
-------------+----------------------------------------------------------------
     sigma_u |   .4272089
     sigma_e |  .20786549
         rho |  .80857291   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. xtfesing ln_wage ttl_exp* union age south c_city not_smsa if year>=85 &
> race==1
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)

                             Robust
     ln_wage        Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]

beta
     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
     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     .120847
         age     .0076781   .0116198     0.66   0.509     -.0150962    .0304525
       south     .0309872    .068533     0.45   0.651      -.103335    .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 with `xtreg`. The homogeneity assumption is not rejected at the 1% level of significance.

Bruno, Magazzini, and Stampini (2020) consider 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.

5 Programs and supplemental materials

To install a snapshot of the corresponding software files as they existed at the time of publication of this article, type

```
. net sj 20-4
. net install st0623      (to install program files, if available)
. net get st0623          (to install ancillary files, if available)
```

6 References

Bruno, R. L., L. Magazzini, and M. Stampini. 2020. Exploiting information from singletons in panel data analysis: A GMM approach. *Economics Letters* 186: Article 108519. https://doi.org/10.1016/j.econlet.2019.07.004.

Bruno, R. L., and M. Stampini. 2009. Joining panel data with cross-sections for efficiency gains. *Giornale degli Economisti e Annali di Economia* 68: 149–173.

Hansen, L. P. 1982. Large sample properties of generalized method of moments estimators. *Econometrica* 50: 1029–1054. https://doi.org/10.2307/1912775.

Verbeek, M. 2004. *A Guide to Modern Econometrics.* 2nd ed. Chichester, UK: Wiley.

Windmeijer, F. 2005. A finite sample correction for the variance of linear efficient two-step GMM estimators. *Journal of Econometrics* 126: 25–51. https://doi.org/10.1016/j.jeconom.2004.02.005.

About the authors

Laura Magazzini is Associate Professor of Econometrics in the Institute of Economics at Sant'Anna School of Advanced Studies, Pisa, Italy (since August 2020). She was previously at the Department of Economics, University of Verona, Italy. She obtained her PhD in 2004 from the Sant'Anna School of Advanced Studies in Pisa, Italy. She graduated in economics and statistics from the University of Florence. During her studies, she was a visiting scholar at Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, and the University of Sydney, Sydney, Australia. Her research interests are centered on 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, and 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 microeconometrics, institutional and comparative economics, labor economics, and innovation from both a macro- and microperspective.

Marco Stampini is Social Protection Lead Specialist at the Inter-American 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. Recently, his studies have focused on conditional cash transfers and long-term care. He holds a master's degree in applied economics from CORIPE Piedmont and a PhD in environmental economics from the Sant'Anna School of Advanced Studies.