The Stata Journal (2015)
15, Number 3, pp. 881–898

Feasible fitting of linear models with N fixed effects

Fernando Rios-Avila
Levy Economics Institute of Bard College
Blithewood-Bard College
Annandale-on-Hudson, NY
friosavi@levy.org

Abstract. In this article, I describe an alternative approach for fitting linear models with multiple high-order fixed effects. The strategy relies on transforming the data before fitting the model. While the approach is computationally intensive, the hardware requirements for the fitting are minimal, allowing for estimation in models with multiple high-order fixed effects for large datasets. I illustrate implementing this approach using the U.S. Census Bureau Current Population Survey data with four fixed effects. I also present a new Stata command, `regxfe`, for implementing this strategy.

Keywords: st0409, regxfe, itercenter, nredound, fixed-effects models, two-step estimation

1 Introduction

With the growing availability of large longitudinal datasets, the use of models with one or more fixed effects has increased. The ability to control for unobserved heterogeneity shared across groups using fixed-effects models appeals to researchers in fields such as economics, sociology, and political science. For example, researchers concerned with the interaction of, for example, firms and workers (Abowd, Kramarz, and Woodcock 2008) or schools, teachers, and students (Harris and Sass 2011) use these models because they allow researchers to control for otherwise unobserved heterogeneity within groups. Depending on the specification of the model, these types of models can be implemented by adding dummy sets that absorb the specific fixed effects.

If the number of groups within a defined category is large, implementing fixed-effects models with dummy sets can be difficult when using standard statistical software. These models often quickly exhaust the computer's memory capacity and thus its ability to manage large matrices of estimated parameters. Furthermore, despite advances in hardware and software, fitting models with multiple high-order fixed effects can be challenging with large datasets.

While linear models with one fixed effect can be fit without including a set of dummy variables as regressors (within estimator and first-difference estimator—see Cameron and Trivedi [2005]), there is no simple solution when there are multiple fixed effects. Much of the literature dealing with estimating these types of models is based on work by Abowd, Kramarz, and Margolis (1999), in which the authors propose various methods for fitting a two-fixed-effects model.[1]

In recent years, many strategies have been developed and implemented for fitting models with one and two high-order fixed effects, with performance varying in terms of systems requirements, computational efficiency, and the estimation of standard errors (McCaffrey et al. 2012).[2] Despite the growing literature on high-order fixed-effects models, the analysis of data with more than two fixed effects is not yet routine. Guimarães and Portugal (2010) presented an algorithm to fit linear models with high-order fixed effects using an iterative conditional regression strategy. This approach was later used in Torres et al. (2013) to fit a three-fixed-effects model. More recently, Gaure (2013) proposed an alternative methodology that generalizes the estimation of high-order fixed effects in linear models, similar to the method I propose. Gaure (2013), however, does not offer a solution for correcting the degrees of freedom when using more than two fixed effects.

In this article, I provide a viable method to make fitting high-order fixed-effects models more accessible, and I demonstrate implementing this methodology using Stata. This methodology is similar to the one suggested in Guimarães and Portugal (2010), but it relies on a different theoretical foundation and implements a more intuitive strategy. In this respect, the methodology presented here is closer to Gaure (2013). I also propose a new algorithm for estimating the exact number of redundant parameters, which produces more precise estimates of the standard errors. I implement this algorithm using Stata, and I demonstrate the procedure using data from the U.S. Census Bureau Current Population Survey (CPS). I also include a Monte Carlo simulation to test the accuracy of the results.

The article is structured as follows. In section 2, I present the base model with one fixed effect. In section 3, I extend the model to a two-fixed-effects model, and I show the generalization for a model with three or more fixed effects in section 4. In section 5, I discuss the estimation of standard deviations. In section 6, I present `regxfe`, the command that implements the proposed algorithm, and I demonstrate using the command with the CPS data and a Monte Carlo simulation to test the accuracy of the algorithm. I conclude in section 7.

2 One-fixed-effect model

Using employer–employee linked data, we consider the basic model with one fixed effect,

$$y_{ijk} = a_i + x_{ijk}\beta + \epsilon_{ijk} \tag{1}$$

where y_{ijk} represents the outcome of person i working at firm j at time k with a total of I individuals and J firms across K periods. In this simple model, we assume that

[1] For details on these methodologies, see Andrews, Schank, and Upward (2006).

[2] McCaffrey et al. (2012) present a review of various commands and strategies created for Stata. Not mentioned in that article are the commands `reg2hdfe` and `reghdfe`, recently released in the Statistical Software Components library.

the outcome of y_{ijk} is a function only of the individual fixed effect a_i and a set H of observed characteristics x_{ijk} that could vary across individual, firm, and time. Finally, we let ϵ_{ijk} be a homoskedastic error term that has a mean of zero and is uncorrelated with a_i and x_{ijk}. Without a loss of generality, we assume that all variables y_{ijk} and x_{ijk} are measured as deviations from their overall sample means.

$$E(\epsilon_{ijk}|x_{ijk}) = 0, \quad E_i(\epsilon_{ijk}) = 0, \quad \text{and} \quad \text{corr}(\epsilon_{ijk}, a_i) = \text{corr}(\epsilon_{ijk}, x_{ijk}) = 0$$

We can fit this model directly, without estimating the actual individual fixed effects, by subtracting the within-person mean from all variables in the model

$$E(y_{ijk}|i = \text{i}) = i_{\overline{y}} = a_i + i_{\overline{x}}\beta \tag{2}$$

where $i_{\overline{y}}$ ($i_{\overline{x}}$) is the within-person i average across all firms and time periods of y (x). Subtracting (2) from (1), we obtain the following transformation of the data (this is called "demeaning" the data):

$$\begin{aligned} y_{ijk} - i_{\overline{y}} &= (x_{ijk} - i_{\overline{x}})\beta + \epsilon_{ijk} \\ \widetilde{y}_{ijk} &= \widetilde{x}_{ijk}\beta + \epsilon_{ijk} \end{aligned} \tag{3}$$

These equations can now be estimated directly using standard ordinary least-squares (OLS) procedures. Note that while the error term ϵ_{ijk} remains unchanged in (3) compared with the original model, the variance–covariance matrix of β (Σ_β) must be corrected to account for the degrees of freedom associated with the unestimated fixed effects (to be discussed in section 5).

3 Two-fixed-effects model

Let's now extend the model to allow for two fixed effects, such that the outcome y is a function of the individual fixed effect a_i and the firm fixed effect b_j, where person i works at time k. Again all variables are measured as deviations from their means.

$$y_{ijk} = a_i + b_j + x_{ijk}\beta + \epsilon_{ijk} \tag{4}$$

Like before, we assume the error term is well behaved and uncorrelated with the explanatory variables, the firm, and the individual fixed effects. Here, if we calculate the within-person average and within-firm average, we obtain

$$E(y_{ijk}|i = \text{i}) = i_{\overline{y}} = a_i + i_{\overline{b}_j} + i_{\overline{x}}\beta \quad \text{and} \tag{5}$$

$$E(y_{ijk}|j = \text{j}) = j_{\overline{y}} = j_{\overline{a}_i} + b_j + j_{\overline{x}}\beta \tag{6}$$

where $i_{\overline{b}_j}$ is the average firm effect from all the firms where person i has ever worked and $j_{\overline{a}_i}$ is the average individual effect from all individuals who have worked for firm j. In both cases, the averages are weighted by the number of times each worker–employer combination is observed.

Note that while $i_{\overline{b}_j}(j_{\overline{a}_i})$ is fixed within individual i (firms j), it still varies across individuals (firms).[3] From (4), we can eliminate part of the impact of the individual and the firm fixed effects. By subtracting the within-group means obtained in (5) and (6), we obtain

$$\begin{aligned} y_{ijk} - i_{\overline{y}} - j_{\overline{y}} &= (x_{ijk} - i_{\overline{x}} - j_{\overline{x}})\beta - j_{\overline{a}_i} - i_{\overline{b}_j} + \epsilon_{ijk} \quad \text{or} \\ \widetilde{y}_{ijk} &= \widetilde{x}_{ijk}\beta - j_{\overline{a}_i} - i_{\overline{b}_j} + \epsilon_{ijk} \end{aligned} \tag{7}$$

While the main components of the individuals and firms fixed effects (a_i and b_j) are eliminated in (7), some heterogeneity remains, and $j_{\overline{a}_i}$ and $i_{\overline{b}_j}$ vary across firms and individuals, respectively. We can eliminate this heterogeneity by continuing to demean the variables in (7) by subtracting the corresponding averages.

$$\begin{aligned} E(y_{ijk} - i_{\overline{y}} - j_{\overline{y}} | i = \mathrm{i}) = i_{\overline{y}} - i_{\overline{y}} - ij_{\overline{y}} &= (i_{\overline{x}} - i_{\overline{x}} - ij_{\overline{x}})\beta - ij_{\overline{a}_i} - i_{\overline{b}_j} \quad \text{or} \\ -ij_{\overline{y}} &= (-ij_{\overline{x}})\beta - ij_{\overline{a}_i} - i_{\overline{b}_j} \end{aligned} \tag{8}$$

$$\begin{aligned} E(y_{ijk} - i_{\overline{y}} - j_{\overline{y}} | j = \mathrm{j}) = j_{\overline{y}} - ji_{\overline{y}} - j_{\overline{y}} &= (j_{\overline{x}} - ji_{\overline{x}} - j_{\overline{x}})\beta - j_{\overline{a}_i} - ji_{\overline{b}_j} \quad \text{or} \\ -ji_{\overline{y}} &= (-ji_{\overline{x}})\beta - j_{\overline{a}_i} - ji_{\overline{b}_j} \end{aligned} \tag{9}$$

Above $ji_{\overline{y}}$ is the within-firm j average of the average outcomes of individuals i, while $ij_{\overline{y}}$ is the within-individual i average of the average outcomes in firm j, both weighted by the number of times each combination is observed. Note that while the expressions $ji_{\overline{y}}$ and $ij_{\overline{y}}$ look similar, they will be the same only in a balanced panel. Subtracting (8) and (9) from (7), we can further reduce the individual and firm heterogeneity and obtain the following expression:

$$\begin{aligned} y_{ijk} - i_{\overline{y}} - j_{\overline{y}} + ij_{\overline{y}} + ji_{\overline{y}} &= (x_{ijk} - i_{\overline{x}} - j_{\overline{x}} + ij_{\overline{x}} + ji_{\overline{x}})\beta + ji_{\overline{a}_i} + ij_{\overline{b}_j} + \epsilon_{ijk} \quad \text{or} \\ \widetilde{y}_{ijk} - i_{\widetilde{y}} - j_{\widetilde{y}} &= (\widetilde{x}_{ijk} - i_{\widetilde{x}} - j_{\widetilde{x}})\beta + ji_{\overline{a}_i} + ij_{\overline{b}_j} + \epsilon_{ijk} \quad \text{or simply} \\ \widetilde{\widetilde{y}}_{ijk} &= \widetilde{\widetilde{x}}_{ijk}\beta + ji_{\overline{a}_i} + ij_{\overline{b}_j} + \epsilon_{ijk} \end{aligned} \tag{10}$$

Once again, while the heterogeneity observed in (7) is no longer present in (10) ($j_{\overline{a}_i}$ and $i_{\overline{b}_j}$), some individual and firm heterogeneity in the form of $ji_{\overline{b}_j}$ and $ij_{\overline{a}_i}$ remains. It can be shown, however, that the variation of the heterogeneity that comes from $ji_{\overline{b}_j}$ and $ij_{\overline{a}_i}$ is lower than what was observed previously in b_j and a_i (see the proof in *Appendix A*). Furthermore, if we continue to demean the variables iteratively, we can achieve a specification where $ji \dots ji_{\overline{b}_j}$ and $ij \dots ij_{\overline{a}_i}$ converge to zero, with their variance also equal to zero.

$$ij \dots ij_{\overline{a}_i} \cong jij \dots ij_{\overline{a}_i} \cong 0 \quad \text{and} \quad ij \dots ij_{\overline{b}_i} \cong jij \dots ij_{\overline{b}_i} \cong 0$$

Thus

$$\mathrm{Var}(ij \dots ij_{\overline{a}_i}) = \mathrm{Var}(ji \dots ji_{\overline{b}_j}) \cong 0$$

[3] The only case in which $i_{\overline{b}_j}$ and $j_{\overline{a}_i}$ are constant across both dimensions (i and j) is when there are the same number of observations for all combinations of i and j, equivalent to a balanced panel.

This effectively eliminates the fixed-effects components from the specification. At this point, the specification can be written as

$$\widetilde{\dddot{\widetilde{y}}}_{ijk} = \widetilde{\dddot{\widetilde{x}}}_{ijk}\beta + \epsilon_{ijk} \tag{11}$$

where

$$\begin{aligned}
\widetilde{\dddot{\widetilde{y}}}_{ijk} &= y_{ijk} - i_{\overline{y}} - j_{\overline{y}} + \cdots - iji \dots\ ji_{\overline{y}} - jij \dots\ ij_{\overline{y}} \\
\widetilde{\dddot{\widetilde{x}}}_{ijk} &= x_{ijk} - i_{\overline{x}} - j_{\overline{x}} + \cdots - iji \dots\ ji_{\overline{x}} - jij \dots\ ij_{\overline{x}}
\end{aligned}$$

We see that (11), like (3), can be directly estimated using standard OLS procedures to obtain unbiased β coefficients.

4 N-fixed-effects model

We can now extend the model to allow for N fixed effects. Assume that the outcomes y are a function of a set of H characteristics x and N fixed effects $n_1, n_2, \dots, n_N$. Each fixed effect n_j contains N_j different groups. To simplify the notation, I have dropped the subscripts because each observation potentially occurs within one group of the N fixed effects. Again, without a loss of generality, we assume that all variables y_{ijk} and x_{ijk} are measured as deviations from their overall sample means.

$$y = x\beta + \Sigma_{i=1}^{N} n_i + \epsilon \tag{12}$$

As before, we can assume the error ϵ is well behaved and uncorrelated with all fixed effects and observed characteristics x. Following the same strategy used for the two-fixed-effects model, we first estimate the means with respect to each fixed-effect group.[4]

$$E(y|i = \mathrm{i}) = i_{\overline{y}} = \Sigma_{j\in N} i_{\overline{n}_j} + i_{\overline{x}}\beta, \quad \text{for all} \quad i = 1, 2 \dots N \tag{13}$$

Subtracting all means in (13) from (12), we start to eliminate the variation coming from N fixed effects. Analogous to what was previously seen, however, some heterogeneity will remain from the averaged fixed effects ($i_{\overline{n}_j}$ for $i \neq j$, and $i, j = 1, \dots, N$).

$$\begin{aligned}
y - \Sigma_{j=1}^{N} j_{\overline{y}} = (x - \Sigma_{j=1}^{N} j_{\overline{x}})\beta - \Sigma_{j\neq 1} 1_{\overline{n}_j} - \Sigma_{j\neq 2} 2_{\overline{n}_j} - \cdots \Sigma_{j\neq N} N_{\overline{n}_j} + \epsilon \quad \text{or} \\
\widetilde{y} = \widetilde{x}\beta - \Sigma_{i=1}^{N}\Sigma_{j\neq i} i_{\overline{n}_j} + \epsilon
\end{aligned} \tag{14}$$

Following a strategy similar to that used for the two-fixed-effects model, we can attempt to eliminate the fixed effects from (14) by obtaining the corresponding averages.

$$k_{\widetilde{y}} = k_{\widetilde{x}}\beta - \Sigma_{i=1}^{N}\Sigma_{j\neq i} ki_{\overline{n}_j} + \epsilon, \quad \text{for all} \quad k = 1, 2 \dots N \tag{15}$$

[4] Note that $i_{\overline{n}_i} = n_i$ and that $ii_{\overline{n}_j} = i_{\overline{n}_j}$.

Using each of the group averages, we proceed to subtract (15) from (14) to eliminate the fixed effects through an iterative demeaning process. Like with the two-fixed-effects model, the transformation will steadily eliminate the influence of the fixed effects from the variables. After multiple iterations, we can obtain a specification similar to (11), which can be estimated using OLS to obtain the unbiased β coefficients as

$$\widetilde{\dddot{\widetilde{y}}} = \widetilde{\dddot{\widetilde{x}}}\beta + \epsilon \tag{16}$$

where

$$\widetilde{\dddot{\widetilde{y}}} = y - \Sigma_{i=1}^{N} i_{\overline{y}} + \Sigma_{i=1}^{N} \Sigma_{k \neq i} ki_{\overline{y}} - \cdots - \Sigma_{i=1}^{N} \Sigma_{k \neq i} \ldots \Sigma_{g \neq h} gh \ldots ki_{\overline{y}}$$

$$\widetilde{\dddot{\widetilde{x}}} = x - \Sigma_{i=1}^{N} i_{\overline{x}} + \Sigma_{i=1}^{N} \Sigma_{k \neq i} ki_{\overline{x}} - \cdots - \Sigma_{i=1}^{N} \Sigma_{k \neq i} \ldots \Sigma_{g \neq h} gh \ldots ki_{\overline{x}}$$

5 Estimating the standard errors

I have discussed how to obtain a specification that allows for the estimation of unbiased β coefficients after accounting for all fixed effects. As seen in (4), (11), and (16), even after the variables have been transformed, the error term remains unchanged and can be used to estimate the variance–covariance matrix.

From (16), after we eliminate the influence of the fixed effects, the corresponding variance–covariance matrix associated with the coefficients β is

$$\mathrm{Var}\left(\widetilde{\beta}\right) = \frac{\Sigma\epsilon^2}{N-k} \times \left(\widetilde{\dddot{\widetilde{x}}}' \widetilde{\dddot{\widetilde{x}}}\right)^{-1}$$

Because the vector of variables $\widetilde{\dddot{\widetilde{x}}}$ is orthogonal to the individual and fixed effects, thus already accounting for the absence of the dummy cross-products in the inverted matrix, the main difference with estimating the original specification is the number of degrees of freedom. If we could fit the original model, we would be required to estimate H parameters for each variable in x and up to $N_1 + N_2 + \cdots N_N - N$ fixed effects (or $I + J - 1$ in the two-fixed-effects case). As noted by Abowd, Kramarz, and Margolis (1999), not all fixed effects can be estimated, because there may not be enough observations to fully identify all fixed effects.

Abowd, Creecy, and Kramarz (2002) presented an algorithm to identify “mobility groups” using two fixed effects (firms and workers). These mobility groups represent the number of parameters among the fixed effects that cannot be identified or estimated in the original model. For models with three or more fixed effects, there is no exact solution for estimating the total number of unidentifiable parameters in the system. Below I propose a modification of the algorithm presented in Abowd, Creecy, and Kramarz (2002) that can be used to find the number of unidentifiable parameters in the model.

To understand the intuitive nature of the strategy, we will start with a model with three fixed effects: n_1, n_2, and n_3. An initial estimate of the number of redundant parameters could be estimated as the sum of all mobility groups (#redundant parameters) of all pair combinations among the fixed effects—it would be the sum of the redundant parameters between (n_1, n_2), (n_1, n_3), and (n_2, n_3). Some mobility groups identified in the first pair of fixed effects may coincide with some of the mobility groups in the second and third pairs. Ignoring this possibility could lead to overstating the total number of redundant parameters. To prevent this, we can subtract the total number of mobility groups between the three fixed effects from the previous expression.

This strategy can be generalized for N fixed effects as follows. Let's assume the model has N fixed effects (12) and call these fixed effects n_1, $n_2, \ldots, n_N$. Let's also define a couple of functions. Let $G_{ij} = G(n_i, n_j)$ be a function that creates a variable that identifies the mobility groups between the variables n_i and n_j using the algorithm presented in Abowd, Creecy, and Kramarz (2002); for three or more variables, let $G_{ijk} = G\{G(n_i, n_j), n_k\}$. Also let $g_{ij} = g(n_i, n_j)$ be a function that identifies the number of mobility groups between the variables n_i and n_j. Similarly, for three or more variables, let $g_{ijk} = g\{G(n_i, n_j), n_k\}$.

The total number of unidentifiable parameters can then be defined as

$$G = \sum_{i<j} g_{ij} + \mathbb{I}_{N\geq 3} \sum_{\eta=3}^{N} \left\{ (-1)^{\eta} \sum_{i<\cdots<h} g_{\underbrace{i\ldots h}_{\eta}} \right\} \quad \text{for} \quad i, j, \ldots, h \leq N$$

where $g_{i\ldots h}$ calculates the number of mobility groups between η variables.

The algorithm presented above suggests a way to create an empirical approximation of the number of unidentified parameters, G. Once G is estimated, the variance–covariance matrix can be corrected using the correct degrees of freedom as follows:

$$\text{Var}\left(\widehat{\beta}\right) = \frac{\Sigma\epsilon^2}{N - K - \Sigma N_i + G} \times \left(\widetilde{\dddot{\widetilde{x}}}' \widetilde{\dddot{\widetilde{x}}} \right)^{-1}$$

6 Estimating the N-fixed-effects model in Stata: The regxfe command

6.1 Syntax

The command `regxfe` implements the algorithms presented above and provides the user with a set of standard statistics comparable with the statistics provided by the standard `regress` and `areg` commands. The syntax for the command is as follows:

```
regxfe depvar [indepvars] [if] [in] [weight], fe(fe_varlist) [xfe(str)
   cluster(varname) robust file(filename) replace mg(#) tolerance(#)
   maxiter(#)]
```

`aweight`s, `fweight`s, and `iweight`s are allowed; see [U] **11.1.6 weight**.

The user can specify up to seven variables to be used as fixed effects and can transform the data before fitting the model. The user can also save the fixed-effects estimates by using the option `xfe(`*str*`)`. The string specified in this option becomes the prefix for the new variables. However, this process increases the computational time required to execute the command.

Using the same assumptions as those used in the `regress` command, `regxfe` allows for the estimation of standard errors assuming homoskedasticity, heteroskedasticity (`robust`), and one-way cluster errors (`cluster(`*varname*`)`). These are different from the cluster standard errors one might obtain using the `xtreg` command because "[both] commands make different assumptions about whether the number of groups increases with the sample size" (StataCorp 2015, 79).

If the user prefers, the transformed data (demeaned), as well as all relevant variables used in the sample, can be saved as a separate file. The user defines the filename using the option `file(`*filename*`)`. If a file with the same name already exists, one can overwrite the file using the `replace` option.

The user can also specify the number of redundant parameters among fixed effects to be used to correct the degrees of freedom in the estimation. This reduces computational time because the program skips the step of internally calculating these parameters.[5]

When convergence is slow, the user can define different levels of `tolerance()` and `maxiter()` criteria. The default is `tolerance(1.192e-07)`, while `maxiter()` allows a maximum of 10,000 iterations. A smaller number of iterations and larger tolerance levels can help reduce processing time but may produce less accurate results. Conversely, lower tolerance levels typically increase the computational time but improve the accuracy of the results.

Even though `regxfe` was originally designed to deal with large datasets, the command may be unable to fit the desired model when it encounters memory limitations. One benefit of this program, however, is that it can be implemented as modules; the user can individually call the routines used in the program, which may mitigate some of the memory constraints often encountered when dealing with large datasets.[6]

[5] The `mg()` option is recommended if the user is running the program by modules when handling particularly large data. See the `nredound` command and illustration in *Appendix B*.

[6] See *Appendix B* for details.

6.2 Illustration

In this section, I implement the algorithm previously described using a sample obtained from the basic CPS monthly survey from 2007 and 2008. The implementation was done using Stata 12 with a Xeon CPU 1.8GHz and 8GB of memory. While the dataset does not provide the richness and complexity typically found in employer–worker linked data or school–teacher data, it illustrates the value of the program for fitting linear models with multiple fixed effects.

We assume a simple wage model,

$$\ln w = \alpha + \beta^{'} X + \texttt{ind} + \texttt{occ} + \texttt{yrm} + \texttt{state} + \epsilon$$

where hourly log wages ($\ln w$) are a function of X [consisting of age (`age`), sex (`sex`), years of education (`yrs_school`), and union status (`union`)], and fixed effects depending on the industry (`ind`) and occupation (`occ`) of the workers, the year × month of the survey (`yrm`), and the state where workers reside (`state`). The error term ϵ is assumed to be homoskedastic and normally distributed. As a benchmark, the model is fit using dummy sets to capture the four fixed effects.

The following is a benchmark regression using the `regress` command:

```
. use cps_sample, clear
. set seed 101
. generate wg=runiform()*50
. set matsize 2000
. regress lnwageh age union yrs_school sex i.yrm i.state i.ind i.occ
      Source |       SS           df       MS      Number of obs   =   165,439
-------------+----------------------------------   F(805, 164633)  =    263.35
       Model |  34737.8417       805  43.1525984   Prob > F        =    0.0000
    Residual |   26976.994   164,633  .163861401   R-squared       =    0.5629
-------------+----------------------------------   Adj R-squared   =    0.5607
       Total |  61714.8358   165,438  .373039059   Root MSE        =     .4048

------------------------------------------------------------------------------
     lnwageh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0055982   .0000822    68.13   0.000     .0054371    .0057592
       union |   .1859473   .0042027    44.24   0.000       .17771    .1941845
  yrs_school |   .0502352    .000516    97.36   0.000     .0492239    .0512465
         sex |   .1420232   .0025609    55.46   0.000     .1370039    .1470426
  (output omitted)
       _cons |   2.912373   .0369547    78.81   0.000     2.839942    2.984803
------------------------------------------------------------------------------
```

We can also fit the same model using the command `regxfe` with the algorithms presented in this article.

The following example is an alternative estimation model using the regxfe command:

```
. regxfe lnwageh age union yrs_school sex, fe(yrm state ind occ)
Transforming the data
.........................
Estimating redundant parameters
...........
Number of redundant parameters: 3
                                                  Number of obs      =   165439
                                                  F_all(  805,164633)=  263.348
                                                  Prob > F_all       =   0.0000
                                                  F_xb (    4,164633)= 4992.053
                                                  Prob > F_xb        =   0.0000
                                                  F_fe (  801,164633)=  124.093
                                                  Prob > F_fe        =   0.0000
                                                  R2-Overall         =   0.5629
                                                  R2-Within          =   0.1082
                                                  # redundant FE     =        3

------------------------------------------------------------------------------
     lnwageh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0055982   .0000822    68.13   0.000     .0054371    .0057592
       union |   .1859473   .0042027    44.24   0.000     .1777101    .1941846
  yrs_school |   .0502352    .000516    97.36   0.000     .0492239    .0512465
         sex |   .1420233   .0025609    55.46   0.000     .1370039    .1470426
       _cons |    1.83969   .0080185   229.43   0.000     1.823974    1.855406
------------------------------------------------------------------------------
```

As we can see, the parameter estimates provided by the regxfe command match those of the benchmark estimation (regress with dummy sets), with differences caused by rounding error. The most important difference between these two commands is the constant estimate.[7] This difference exists because with the regress command, the constant captures the value of the omitted fixed-effects categories. In contrast, when we run regxfe, the fixed effects are calculated as deviations from zero, and the constant is not affected by any omitted fixed-effects category.

Along with standard statistics, regxfe reports the overall F statistic (F_all), the F statistic corresponding to the explanatory variables (F_xb), and the F statistic for the joint test of all fixed effects equal to zero (F_fe). However, the last two statistics are not available if the model is fit using the robust or cluster() option. regxfe also reports the R^2 for the full model (including fixed effects) and the within-R^2, which is associated with the "goodness of fit" of all parameters after taking into account the influence of the fixed effects. Finally, regxfe also provides the number of redundant parameters within the N fixed effects.[8] For this example, the algorithm takes 22 iterations before it converges using the default float precision, taking a total of 59 seconds.

As shown in figure 1, the parameters are very close to their true values after less than 10 iterations. This should be taken into consideration, with the understanding

[7] In *Appendix C*, we provide results for similar models using weights and using the robust and cluster() options.

[8] This parameter does not take into account the loss of one parameter due to the constant.

that processing time could increase rapidly for larger datasets and for more complex specifications.

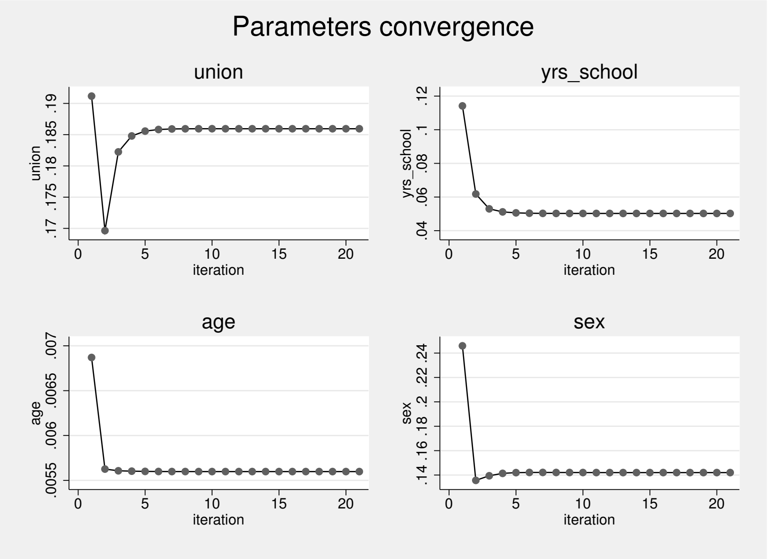

Figure 1. Parameters convergence by iteration

6.3 Monte Carlo simulation

The example presented above is a useful illustration of the capabilities of the `regxfe` command for fitting linear models with multiple fixed effects. However, the relatively simple structure of the fixed effects may mean that the correspondence of the point estimates and standard deviations could be random. In this section, I present a Monte Carlo simulation to better assess the accuracy of the proposed method.

Data structure: Data-generation process

To proceed with the simulation analysis, we must first establish the underlying data structure that captures particular features of real data, specifically a structure that mimics a complex structure of fixed-effects interrelations.[9]

Here we create datasets with four explanatory variables and four fixed effects. The fixed-effects groups are generated so that there are m_1 mobility groups between the first and second fixed effects, m_2 mobility groups between the second and third fixed effects, and at least two mobility groups between the first, second, and fourth fixed effects. The values associated with the impact of each fixed effect are generated using a standard

[9] The program used to create the simulated data is available upon request.

normal distribution. To create the explanatory variables, we generate the latent values using a multivariate standard normal distribution with correlation matrix $\mathbf{\Sigma}$.

$$X \in (x_1, x_2, x_3, x_4) \sim N(0, \mathbf{\Sigma})$$

To model the presence of correlation between the fixed effects and the explanatory variables, we define the final values of X as

$$x_i = x_i + \delta_{i1}FE1 + \delta_{i2}FE2 + \delta_{i3}FE3 + \delta_{i4}FE4 \quad \text{for} \quad i = 1, 2, 3, \text{ and } 4$$

where the values for δ_{ij} are determined using a uniform distribution between -1 and 1. Finally, the dependent variable is defined as

$$y = 1 + x_1 + x_2 + x_3 + x_4 + FE1 + FE2 + FE3 + FE4 + \epsilon$$

where ϵ follows a normal distribution with mean 0 and standard deviation 3.

Simulation results

Based on the data structure described above, we generate 5 sets of 100 datasets containing 2.5K, 7.5K, 15K, 50K, and 250K observations.

For each dataset, the model implied above is fit with OLS with dummy variables to control for the fixed effects (`regress`) and with `regxfe` with default options. Using the `regress` command as a benchmark, table 1 presents the average and maximum absolute difference between the `regress` and the `regxfe` estimates.

Table 1. Summary statistics: Absolute difference between `regress` and `regxfe` estimates

	2.5K Obs		10K Obs		50K Obs		250K Obs	
	Mean	Max	Mean	Max	Mean	Max	Mean	Max
`b1`	1.62E–06	1.80E–05	8.32E–10	1.41E–08	2.7E–07	1.86E–05	1.56E–06	5.86E–05
`se(b1)`	3.22E–06	3.47E–05	1.50E–09	1.92E–08	4.59E–07	3.29E–05	2.57E–06	0.000131
`b2`	3.84E–06	4.04E–05	1.87E–09	2.33E–08	4.55E–07	3.11E–05	3.45E–06	0.000164
`se(b2)`	5.01E–06	5.07E–05	2.52E–09	3.72E–08	3.57E–07	2.14E–05	4.57E–06	0.000211
`b3`	1.53E–07	1.50E–06	3.17E–11	4.80E–10	3.68E–09	3.60E–07	7.66E–09	4.71E–07
`se(b3)`	3.77E–07	3.11E–06	7.79E–11	1.16E–09	5.84E–09	5.73E–07	1.42E–08	5.79E–07
`b4`	4.46E–07	3.98E–06	9.96E–11	1.41E–09	6.23E–09	5.98E–07	2.22E–08	9.34E–07
`se(b4)`	6.10E–07	4.89E–06	1.33E–10	1.95E–09	1.17E–08	1.06E–06	2.98E–08	1.37E–06
`df_m`	0	0	0	0	0	0	0	0
`df_r`	0	0	0	0	0	0	0	0

Note: For each sample size, 100 datasets were created following the data structure described above. `regress` is used by including dummy variables. The `regxfe` is used with the default options.

As seen in table 1, the algorithm used to calculate the number of degrees of freedom provides the same number obtained using the `regress` command. Although using the standard tolerance level (`epsfloat()`) does not provide an exact match on the point estimates of the model when comparing both command's estimates, the results are within rounding error. The maximum absolute difference observed within the 100 simulated samples is 0.000164 for parameter `b2` and 0.000211 for parameter `se(b2)`, both in the 250K observation samples.

7 Conclusions

In this article, I present a new command, `regxfe`, that can be used to fit linear models with N fixed effects. The command implements an algorithm that uses an intuitive strategy involving an iterative demeaning process, which is already used for fitting models with one fixed effect (`areg`). I also propose an algorithm to estimate the total number of redundant parameters, which allows for more precise estimates of the standard errors. The results from the Monte Carlo simulation show that point estimates obtained using `regxfe` closely match those using `regress`, matching within rounding error, with an exact match on the estimation of degrees of freedom in the models.

While this strategy is computationally intensive, it provides some advantages compared with other competing strategies currently available in Stata. First, it allows for the use of weights when fitting fixed-effects models; second, the new proposed algorithm provides an exact estimate of the degrees of freedom of the model; and third, the modules of this algorithm can be used individually, which can mean a more efficient use of computer memory when dealing with large datasets and limited computer resources. For instance, Hotchkiss, Quispe-Agnoli, and Rios-Avila (2015) used a previous version of this methodology to fit a fixed-effects model with 4 fixed effects (3,376,102 workers, 93,021 firms, 40 quarters, and 159 counties)—a model that could not be fit using other available strategies.

Computational limitations will diminish with time if current trends in hardware development continue. However, the desire to use ever larger datasets and models with larger numbers of fixed effects may leave us with the same computational constraints. We cannot control such events. However, the flexibility offered with this program, I hope, might offer the research community some valuable tools for fitting models with N fixed effects.

8 Acknowledgments

I thank Julie L. Hotchkiss, Thomas Masterson, and Tim Sass for their helpful comments in completing this article. I also give special thanks to Gustavo Canavire-Bacarreza for his help with the simulation programming and to an anonymous referee for comments and insights on the article and the program implementation.

9 References

Abowd, J. M., R. H. Creecy, and F. Kramarz. 2002. Computing person and firm effects using linked longitudinal employer–employee data. Technical Paper No. TP-2002-06, Center for Economic Studies, U.S. Census Bureau. http://www2.census.gov/ces/tp/tp-2002-06.pdf.

Abowd, J. M., F. Kramarz, and D. N. Margolis. 1999. High wage workers and high wage firms. *Econometrica* 67: 251–333.

Abowd, J. M., F. Kramarz, and S. Woodcock. 2008. Econometric analyses of linked employer—employee data. In *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, ed. L. Mátyás and P. Sevestre, 3rd ed., 727–760. Berlin: Springer.

Andrews, M., T. Schank, and R. Upward. 2006. Practical fixed-effects estimation methods for the three-way error-components model. *Stata Journal* 6: 461–481.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

Gaure, S. 2013. OLS with multiple high dimensional category variables. *Computational Statistics and Data Analysis* 66: 8–18.

Guimarães, P., and P. Portugal. 2010. A simple feasible procedure to fit models with high-dimensional fixed effects. *Stata Journal* 10: 628–649.

Harris, D. N., and T. R. Sass. 2011. Teacher training, teacher quality and student achievement. *Journal of Public Economics* 95: 798–812.

Hotchkiss, J. L., M. Quispe-Agnoli, and F. Rios-Avila. 2015. The wage impact of undocumented workers: Evidence from administrative data. *Southern Economic Journal* 81: 874–906.

McCaffrey, D. F., J. R. Lockwood, K. Mihaly, and T. R. Sass. 2012. A review of Stata commands for fixed-effects estimation in normal linear models. *Stata Journal* 12: 406–432.

StataCorp. 2015. *Stata 14 Base Reference Manual*. College Station, TX: Stata Press.

Torres, S., P. Portugal, J. T. Addison, and P. Guimarães. 2013. The sources of wage variation: A three-way high-dimensional fixed effects regression model. IZA Discussion Paper No. 7276, Institute for the Study of Labor (IZA). http://ftp.iza.org/dp7276.pdf.

About the author

Fernando Rios-Avila is a research scholar working on the Levy Institute Measure of Economic Well-Being under the Distribution of Income and Wealth program. His research interests include labor economics, applied microeconomics, development economics, and poverty and inequality.

Appendix A

Let y_i be a variable with an overall mean $\overline{y}$ and variance σ_y^2. Without loss of generality, assume that the ith means of y_i ($i_{\overline{y}}$) are all different from each other; that is, $\sigma_{i_{\overline{y}}}^2 \neq 0$.

The variance of variable y can then be written as

$$\sigma_y^2 = \text{var}(y_i) = E\left\{(y_i - \overline{y})^2\right\}$$

Maintaining the equality on the expression, we can add and subtract the ith mean to the expression in parentheses, obtaining the alternative variance expression

$$\sigma_y^2 = E\left\{(y_i - i_{\overline{y}} + i_{\overline{y}} - \overline{y})^2\right\}$$

Expanding this expression, we obtain

$$\begin{aligned}
\sigma_y^2 &= E\left\{(y_i - i_{\overline{y}})^2 + (i_{\overline{y}} - \overline{y})^2 + 2(y_i - i_{\overline{y}})\,(i_{\overline{y}} - \overline{y})\right\}\\
\sigma_y^2 &= E\left\{(y_i - i_{\overline{y}})^2\right\} + E\left\{(i_{\overline{y}} - \overline{y})^2\right\} + 2E\left\{(y_i - i_{\overline{y}})(i_{\overline{y}} - \overline{y})\right\}\\
\sigma_y^2 &= E\left\{(y_i - i_{\overline{y}})^2\right\} + E\left\{(i_{\overline{y}} - \overline{y})^2\right\} + 2\left\{E(y_i i_{\overline{y}} - i_{\overline{y}}^2)\right\}
\end{aligned}$$

Using iterative expectations, we see that the expression is equal to zero.

$$E(y_i i_{\overline{y}} - i_{\overline{y}}^2) = E\left\{E(y_i i_{\overline{y}} - i_{\overline{y}}^2 | i = \text{i})\right\} = 0$$

Thus

$$\sigma_y^2 = E\left\{(y_i - i_{\overline{y}})^2\right\} + E\left\{(i_{\overline{y}} - \overline{y})^2\right\} \rightarrow \sigma_y^2 = \sigma_{y-i_{\overline{y}}}^2 + \sigma_{i_{\overline{y}}}^2$$

Finally, the overall variance of y can be decomposed into two components: one corresponding to the within-variation, $\sigma_{y-i_{\overline{y}}}^2$, and one corresponding to the across-variation, $\sigma_{i_{\overline{y}}}^2$. Given that all variances must be positive by construction, the variance of the demeaned data and the within-group means must be smaller than that of the original data.

We can further decompose the across-individual variance $\sigma_{i_{\overline{y}}}^2$ with respect to an alternative subgroup. If, for example, we decompose in relation to the groups j, it follows that

$$\sigma_{i_{\overline{y}}}^2 = \sigma_{i_{\overline{y}}-ji_{\overline{y}}}^2 + \sigma_{ji_{\overline{y}}}^2$$

Using the same strategy with iterative decompositions, we see that the variance of subsequent transformations tends to zero.

$$\sigma_y^2 > \sigma_{i_{\overline{y}}}^2 > \sigma_{ji_{\overline{y}}}^2 > \cdots > \sigma_{ij\ldots\, ji_{\overline{y}}}^2 \cong 0$$

Appendix B

While this program was originally created to fit linear models with multiple effects using large datasets, it may fail to fit such models when Stata exhausts the available memory.

In such situations, the modules used in this program can be individually executed before fitting the final linear model. Here I briefly explain the use of two commands that are internally called from `regxfe`, which can be directly used for fitting this type of model.

The itercenter command

`itercenter` *varlist* [*if*] [*in*] [*weight*], `fe(`*fe_varlist*`)` [`tolerance(`*#*`)` `maxiter(`*#*`)` `mean` `replace` `xfe(`*str*`)`]

This command implements the iterative demeaning process of all variables in *varlist* for groups defined by the categorical variables listed in *fe_varlist*. Specifically, it uses the variables listed in *fe_varlist*. The command will replace the original variables with their demeaned transformations. You can specify a `tolerance()` level other than the default (1–e7) as well as the maximum number of iterations (`maxiter(10000)`). If the option `mean` is specified, the transformed variable will preserve the overall mean. The transformed variable will be stored as double format. The option `mean` is used by default in the `regxfe` command to recover the constant.

The nredound command

`nredound` *varlist* [*if*] [*in*]

This command calculates the exact number of redundant parameters given a list of variables, *varlist*, that identify the fixed effects in the dataset. The number of redundant parameters is saved in the local `e(M)`. This can be used to provide `regxfe` with the number of redundant parameters given the identified fixed-effects variables.

Illustration

Below the dependent and independent variables are demeaned in two steps. The same process can be done using partition datasets. This can be used for parallel processing within the same machine or within multiple machines.

Calculating the redundant parameters requires specifying only the *varlist* of the fixed effects.

Special attention is required to ensure that all steps are processed using the same sample.

```
. itercenter lnwageh age union, fe(yrm state ind occ) mean replace
......................
. itercenter union yrs_school sex, fe(yrm state ind occ) mean replace
..........................
. nredound yrm state ind occ
...........
Number of redundant parameters: 3
. local x=e(M)
. regxfe lnwageh age union yrs_school sex, fe(yrm state ind occ) maxiter(0)
> mg(`x´)
Transforming the data
                                                  Number of obs      =   165439
                                                  F_all(  805,164633)=   24.805
                                                  Prob > F_all       =   0.0000
                                                  F_xb (    4,164633)= 4992.053
                                                  Prob > F_xb        =   0.0000
                                                  F_fe (  801,164633)=    0.000
                                                  Prob > F_fe        =   1.0000
                                                  R2-Overall         =   0.1082
                                                  R2-Within          =   0.1082
                                                  # redundant FE     =        3
------------------------------------------------------------------------------
     lnwageh |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0055982   .0000822    68.13   0.000     .0054371    .0057592
       union |   .1859473   .0042027    44.24   0.000     .1777101    .1941846
  yrs_school |   .0502352    .000516    97.36   0.000     .0492239    .0512465
         sex |   .1420233   .0025609    55.46   0.000     .1370039    .1470426
       _cons |    1.83969   .0080185   229.43   0.000     1.823974    1.855406
------------------------------------------------------------------------------
```

In this case, because we are using the transformed data directly, the `F_all` statistic and `F_fe` statistic are not the correct ones; however, the `F_xb` and the standard deviations match those of the full command's syntax. Similarly, `R2-Overall` does not reflect the overall fit of the model.

Appendix C

Table 2. Alternative specifications of the wage-equation model: `regress` command

	Analytic weights	Robust errors	Cluster errors
`age`	0.005651	0.005598	0.005598
	(0.00008)	(0.00008)	(0.00076)
`union`	0.183397	0.185947	0.185947
	(0.00419)	(0.00374)	(0.00488)
`yrs_school`	0.050034	0.050235	0.050235
	(0.00052)	(0.00056)	(0.00100)
`sex`	0.141389	0.142023	0.142023
	(0.00256)	(0.00276)	(0.00692)
`_cons`	2.946639	2.912373	2.912373
	(0.03609)	(0.03904)	(0.05529)

Note: Weights were created using a uniform randomly generated number between 0–50. The `cluster()` option uses `age` as a cluster variable. All models include dummy variables to account for the year–month, state, industry, and occupation fixed effect.

Table 3. Alternative specifications of the wage-equation model: `regxfe` command

	Analytic weights	Robust errors	Cluster errors
`age`	0.005651	0.005598	0.005598
	(0.00008)	(0.00008)	(0.00076)
`union`	0.183397	0.185948	0.185948
	(0.00419)	(0.00374)	(0.00488)
`yrs_school`	0.050034	0.050235	0.050235
	(0.00052)	(0.00056)	(0.00100)
`sex`	0.141389	0.142024	0.142024
	(0.00256)	(0.00276)	(0.00692)
`_cons`	1.841232	1.83969	1.83969
	(0.00802)	(0.00875)	(0.03823)

Note: Weights were created using a uniform randomly generated number between 0–50. The `cluster()` option uses `age` as a cluster variable.