An Iterative Approach to Estimation with Multiple High-Dimensional Fixed Effects

Siyi Luo[a*], Wenjia Zhu[b], Randall P. Ellis[a]

August 21, 2017

[a] Department of Economics, Boston University
[b] Department of Health Care Policy, Harvard Medical School

Abstract

We develop a new estimation algorithm for models with multiple high-dimensional fixed effects and unbalanced panels. By Frisch-Waugh-Lovell Theorem, our algorithm absorbs fixed effects iteratively until they are asymptotically eliminated. Monte Carlo simulations show that our approach matches results from estimation with fixed effect dummies. Applying the algorithm to US employer-based health insurance data, we analyze health care utilization of 63 million individual-months with fixed effects for 1.4 million individuals, 150,000 primary care physicians, 3,000 counties, 465 employer*year*single/family coverage types and 47 months. We find that narrow network plans reduce the probabilities of monthly visits relative to preferred provider organizations.

Keywords: multiple high-dimensional fixed effects, big data, iterative algorithm, Monte Carlo simulations, health care utilization

JEL codes: I11, G22, I13

Acknowledgements: We are grateful to Iván Fernández-Val, Coady Wing, and participants at International Health Economics Association (Boston), American Society of Health Economists (U. Penn), Annual Health Econometrics Workshop (Honolulu), and Boston University for useful comments and insights. All remaining errors belong to the authors.

[*] Correspondence to: Siyi Luo, Boston University, Boston, MA, USA. E-mail: syluo@bu.edu

1. Introduction

The simplest way to estimate a two-way fixed effect model is to include fixed effects as dummy variables and obtain the least squares dummy variable (LSDV) estimator. When the numbers of levels for both fixed effects are small, using the LSDV is straightforward. When only one of the fixed effects has a large number of levels (i.e., the fixed effect is high dimensional), it is often feasible to include the other fixed effect as dummies. This leaves one high-dimensional fixed effect to absorb and we can apply the usual method of one-way fixed effect model after absorption. In both cases, the LSDV approach will work well theoretically regardless of whether data are balanced or not.

The LSDV method, however, can become computationally infeasible as sample sizes and the numbers of high-dimensional fixed effects increase. One alternative is to estimate transformed models in which fixed effects are eliminated. Balázsi, Mátyás, and Wansbeek (2015) show that in a simple model with two fixed effects and balanced data, the within transformation has a straightforward formula. For models with more than two fixed effects and under common data issues such as unbalanced data, transformation can become intractable. Another transformation for multiple high-dimensional fixed effects is sequential demeaning over fixed effect indices. Such static transformation can completely eliminate fixed effects with balanced panels but will in general fail with unbalanced panels.

In this paper, we propose an alternative approach to transform models featuring large, unbalanced datasets and multiple high-dimensional fixed effects. As opposed to the within transformation which absorbs fixed effects in one step, or the sequential demeaning that is

operated once for all fixed effects, our method demeans variables with respect to each one of the fixed effects sequentially and iteratively. We propose an assumption under which the algorithm converges in the sense that the remaining fixed effects are asymptotically eliminated and the estimator obtained from each iteration converges to the LSDV estimator. Also, this method can be generalized to more complicated models such as those containing more than two high-dimensional fixed effects and instrumental variables without increasing the complexity of the algorithm. Finally, we implement this method in SAS that is particularly capable of handling large data sets.

Guimaraes and Portugal (2010) develop an alternative algorithm that uses the iteration and convergence implementation of least squares estimation with condensed fixed effect variables to reduce the number of explanatory variables. It starts with any initial values of fixed effects and iterates to continuously correct these values by averaging out the fixed effects from residuals. After convergence of the estimates, the fixed effects remain identifiable. An efficient GP algorithm has been programmed as a user build-in function in Stata (Correia 2015) called *reghdfe*, which we use as a benchmark in Monte Carlo simulations. Another alternative algorithm for two-way high-dimensional fixed effect models is from Somaini and Wolak (2016). This algorithm utilizes the common within transformation to absorb one of the fixed effects, and stores the inverse matrix partitioned on the dummies of the other fixed effect in a memory efficient way. Their method is restricted to dealing with two high-dimensional fixed effects.

Table 1 compares our algorithm (denoted as TSLSFECLUS) to other existing programs in Stata and SAS. Our algorithm is able to accommodate all the data features being studied

including multiple high-dimensional fixed effects, 2SLS, clustered standard errors, and large data sets, while the rest of the programs are able to address some of them. *reghdfe* is the program closest to ours, but tends to fail on extremely large datasets due to its heavy use of memory during execution.

Our algorithm involves the following: (1) Absorb fixed effects sequentially from all dependent and explanatory (including instrumental) variables; (2) estimate the model using the demeaned variables; (3) repeat iteratively until the estimates of parameters converge.

We perform Monte Carlo simulations to evaluate the performance of our algorithm. We vary models according to the number of missing observations, dependence of the control variable on one of the fixed effects, dependence of the instrument on one of the fixed effects, extent of endogeneity, range of time fixed effects, model noisiness, and whether errors are clustered or not. Our results match well with those from estimation with fixed effect dummies in all the variations considered.

The proposed algorithm is applied to US employer-based health insurance market data to examine how health plan types affect health care utilization. Our analysis sample, described more fully in Ellis and Zhu (2016), contains about 63 million observations from which we remove fixed effects for 1.4 million individuals, 3,000 counties, 150,000 primary care doctors, 465 employer*year*single/family coverage, and 47 months to predict plan type effects on monthly health care utilization. By simultaneously controlling for all fixed effects, the identification comes from consumers' movement between health plan types. We use propensity scores for household choice of each health plan type as the instrumental

variables to control for endogenous plan choice and correct the standard errors to cluster at the employer level. Our estimates show that the breadth of provider networks dominates cost sharing in influencing consumers' decision to seek care.

The rest of the paper is organized as follows. Section 2 describes our iterative algorithm in a two-way fixed effect model framework. We present two theorems showing that under the proposed assumption, our algorithm generates estimates that are equivalent to those from the simple LSDV model. In Section 3, we conduct Monte Carlo simulations to evaluate the validity and properties of our algorithm. We then show in Section 4 that our algorithm is feasible to estimate a model of health care utilization on a real data set that requires controlling simultaneously for patients, providers and counties, each of high dimension. Finally, Section 5 concludes and discusses further research directions.

2. An Iterative Estimation Algorithm

Consider a simple linear two-way fixed effect model:

$$y_{it} = x_{it}'\beta + \alpha_i + \theta_t + u_{it} \tag{1}$$

$$\forall i \in N_t \subseteq \mathcal{N} = \{1, 2, \dots, N\}$$

$$\text{or } \forall t \in T_i \subseteq \mathcal{T} = \{1,2, \dots, T\}$$

When $N_t \equiv \mathcal{N}, \forall t$ or $T_i \equiv \mathcal{T}, \forall i$, the data is a balanced panel. Otherwise, there are missing observations and the panel is unbalanced. Without any loss of generality, assume $\mathcal{N} \geq \mathcal{T}$. The higher-dimensional fixed effect, i.e. α_i, is always absorbed first.

The idea of our algorithm is to sequentially absorb fixed effects repeatedly until the model converges. Specifically, we employ the following three-step procedure.

Step 1: Demean all dependent and independent variables by each one of the fixed effects sequentially and estimate the model with the demeaned variables. Denote as the 1st iteration.

Step 2: Demean the standardized variables from the previous iteration and estimate the model with the demeaned variables.

Step 3: Repeat Step 2 iteratively until the estimates converge.

Using Model (1), our algorithm can be laid out as the following.

1st iteration:

1) Demean y_{it} and x_{it} over i

$$y_{i\cdot} = x_{i\cdot}'\beta + \alpha_i + \frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t + u_{i\cdot}$$

$$\tilde{y}_{it} \equiv y_{it} - y_{i\cdot} = \tilde{x}_{it}'\beta + \theta_t - \frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t + \tilde{u}_{it}$$

2) Demean the resulting $\tilde{y}_{it}$ and $\tilde{x}_{it}$ over t

$$\tilde{y}_{\cdot t} = \tilde{x}_{\cdot t}'\beta + \theta_t - \frac{1}{\|N_t\|}\sum_{i\in N_t}\frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t + \tilde{u}_{\cdot t}$$

$$\tilde{y}_{it}^{(1)} \equiv \tilde{y}_{it} - \tilde{y}_{\cdot t} = \tilde{x}_{it}^{(1)\prime}\beta - \frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t + \frac{1}{\|N_t\|}\sum_{i\in N_t}\frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t + \tilde{u}_{it}^{(1)}$$

$$= \tilde{x}_{it}^{(1)\prime}\beta + \alpha_i^{(1)} + \theta_t^{(1)} + \tilde{u}_{it}^{(1)}$$

where $\theta_t^{(1)}$ and $\alpha_i^{(1)}$ are the remaining fixed effects after 1st iteration defined as:

$$\begin{cases} \theta_t^{(1)} \equiv \frac{1}{\|N_t\|}\sum_{i\in N_t}\frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t \\ \alpha_i^{(1)} \equiv -\frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t \end{cases}$$

Similarly, the model after (k+1)th iteration can be written as:

$$\tilde{y}_{it}^{(k+1)} \equiv \tilde{y}_{it}^{(k)} - \tilde{y}_{\cdot t}^{(k)} = \tilde{x}_{it}^{(k+1)\prime}\beta + \alpha_i^{(k+1)} + \theta_t^{(k+1)} + \tilde{u}_{it}^{(k+1)}$$

and

$$\begin{cases} \theta_t^{(k+1)} \equiv \frac{1}{\|N_t\|}\sum_{i\in N_t}\frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t^{(k)} \\ \alpha_i^{(k+1)} \equiv -\frac{1}{\|T_i\|}\sum_{t\in T_i}\theta_t^{(k)} \end{cases} \quad (2)$$

ASSUMPTION 1 *Quasi-Balance*: For any two time periods *s* and *t*, there exists an individual who is observed in both periods.

$\forall s, t \in \{1,2,\ldots,T\}, \exists i \text{ s.t. } i \in N_t \text{ and } i \in N_s$

Assumption 1 restricts the extent of unbalancedness of a dataset, although it is in fact a relatively loose condition which could be commonly observed in most empirical datasets.

This assumption is specific to the order of fixed effects being absorbed, i.e., individual fixed effect absorbed first followed by time fixed effect. An example of when Assumption 1 fails is when there are two periods in which the pools of individuals are completely different.[1]

THEOREM 1: For model (1), if the dataset satisfies Assumption 1, then starting from any initial value of $\theta^{(0)} \equiv \theta$, the remaining fixed effects converge to a constant vector, i.e. $\theta_t^{(k)} \to C \ \forall t$, as $k \to \infty$. *(Proof in Appendix A.)*

By Theorem 1, $\theta_t^{(k)} \to C \ \forall t$ and we can easily derive from equation (2) that $\alpha_i^{(k)} \to -C \ \forall i$. So the remaining individual and time fixed effects will be cancelled out with each other upon convergence. In other words, by iterating the sequential absorption, fixed effects are eliminated asymptotically.

THEOREM 2: For model (1) under Assumption 1, the OLS/2SLS estimator $\hat{\beta}^{(k)}$ from each iteration converges to the LSDV estimator of β. *(Proof in Appendix A.)*

COROLLARY 1: If the OLS/2SLS estimator $\hat{\beta}^{(k)}$ converges to the LSDV estimator, the remaining fixed effects converge. *(Proof in Appendix A.)*

Ideally, once the remaining fixed effects converge, we can obtain an unbiased and consistent estimator under regulatory assumptions. However, in practice, convergence of remaining fixed effects could not be observed explicitly. From Theorem 2 and Corollary 1,

[1] When Assumption 1 fails, the time fixed effects cannot be identified because individual fixed effects are nested within these two-period time fixed effects. In addition, if Assumption 1 fails for more than one pair of time periods, then the singularity problem will cause LSDV to fail as well.

it is equivalent to checking the convergence on estimates from the demeaned model in each iteration $\{\hat{\beta}^{(1)}, \hat{\beta}^{(2)}, \dots, \hat{\beta}^{(k)}, \dots, \hat{\beta}^{(\infty)}\}$. For a detailed description of the program implementation, see Appendix B. With endogeneity, we need to modify model (1) into a two-stage linear model and also demean any instrumental variables in Step 1 and 2. However, with these minimal changes the two theorems in Section 2 remain valid, hence our algorithm applies to linear models with endogeneity.

3. Monte Carlo Simulations

3.1. Pseudo Data Generating Process

To examine the performance and convergence properties of our algorithm, we generate pseudo data sets according to model (1). In particular, we allow linear dependence between the fixed effects and the control variable *x*. We also allow the flexibility to include clustered errors that introduce correlation of errors within clusters.

Denote N = number of individuals; T = number of time periods; ρ_{XFEi} = dependence of control variable on individual fixed effect; ρ_{XFEt} = dependence of control variable on time fixed effect; ρ_{ZFEi} = dependence of the potential instrumental variable on individual fixed effect; ρ_{ZFEt} = dependence of the potential instrumental on time fixed effect; M = number of missing observations.

$$y_{it} = 2 * x_{it} + \alpha_i + \theta_t + u_{it}$$

$$x_{it} = z_{it} + \rho_{XFEi} * \alpha_i + \rho_{XFEt} * \theta_t + v_{it}$$

$$z_{it} = z_{it}^{*} + \rho_{ZFEi} * \alpha_i + \rho_{ZFEt} * \theta_t$$

where

$$\alpha_i, z_{it}^{*} \sim U[0,10], \theta_t \sim U[0, \bar{\theta}]$$

$$\begin{pmatrix} u_{it} \\ v_{it} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} \sigma_u^2 & \sigma_u\rho_{uv} \\ \sigma_u\rho_{uv} & 1 \end{pmatrix}\right]$$

$$i = 1, 2, \dots, N, \qquad t = 1, 2, \dots, T$$

When the control variable x_{it} is exogenous, ρ_{uv} equals 0.

To build in clustering of standard errors, we construct errors $\{u_{it}\}_{i,t}$ assuming without loss of generality that errors are clustered within individuals over time:

$$For\ i = 1, 2, \dots, N$$

$$u_i \sim N(0, \sigma_u^2)$$

$$Construct\ u_{it} = \lambda^{t-1} u_i, \lambda \neq 0,1, for\ t = 1, 2, \dots, T$$

where λ governs the serial correlation of errors within individuals and is fixed at 0.5 in all the simulations.[2] When standard errors are clustered, we express the error terms in the model as:

$$\begin{pmatrix} u_{it} \\ v_{it} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} \lambda^{t-1}\sigma_u^2 & \sqrt{\lambda^{t-1}}\sigma_u\rho_{uv} \\ \sqrt{\lambda^{t-1}}\sigma_u\rho_{uv} & 1 \end{pmatrix}\right]$$

We construct unbalanced data by randomly selecting *M* observations to drop from the balanced data, satisfying the Quasi-Balance assumption.

[2] Note that λ cannot be equal to 0 or 1, because otherwise there would be no variation in errors within individuals and u_{it} would be completely absorbed in the same manner as the individual fixed effect α_i, in which case no random errors would be left in the model.

3.2. Variations in Parameters

Our simulation model according to (1) defines 7 random variables $\{y, x, z, \alpha, \theta, u, v\}$ and 10 parameters $\{N, T, M, \rho_{ZFEi}, \rho_{ZFEt}, \rho_{XFEi}, \rho_{XFEt}, \rho_{uv}, \bar{\theta}, \sigma_u\}$. We fix $N = T = 100$, $\rho_{XFEi} = 0.2$ and $\rho_{ZFEi} = 0$, while allowing the rest of parameters to vary.[3] We focus on unbalanced models with two-way fixed effects for both OLS and 2SLS.[4] For each model, we first run 100 simulations and estimate the model using our iteration procedure, and then compare it with the LSDV estimate or equivalently the estimate from the optimal within transformation output by Stata.[5]

We conduct simulations by varying one or a pair of parameters at a time while keeping other parameters fixed at the baseline values, as shown in Table 2. The baseline is $\{M, \rho_{ZFEt}, \rho_{XFEt}, \rho_{uv}, \bar{\theta}, \sigma_u\}$= {5000, 0, 80, 0, 100, 10} for OLS and $\{M, \rho_{ZFEt}, \rho_{XFEt}, \rho_{uv}, \bar{\theta}, \sigma_u\}$ = {5000, 40, 40, 0.6, 100, 10} for 2SLS models. We explore variations along (1) number of missing observations M={5000, 6000, 2000}, (2) relative importance of time fixed effect in explanatory variables $(\rho_{XFEt}, \bar{\theta})$= {(80, 100), (40, 100), (0, 100), (-80, 100), (80, 10)} for OLS and $(\rho_{ZFEt}, \rho_{XFEt}, \bar{\theta})$= {(40, 40, 100), (0, 80, 100), (80, 0, 100), (40, -40, 100), (60, -20, 100), (20, -20, 100), (40, 40, 10)} for 2SLS, (3) model noisiness σ_u= {10, 100}, and (4) correlation between the endogenous variable and the error term, ρ_{uv}= {0.6, 0.2} for 2SLS models only. Finally, we examine OLS models with and without clustered standard errors.

[3] The dependence of x and z on individual fixed effect does not affect simulation results since individual fixed effect is completely absorbed in the first iteration, based on the order of absorption. Therefore we keep them fixed in all the simulations.
[4] Consistent with the analytical model, we find that for balanced data, convergence always happens after the first iteration.
[5] We use Stata's built-in programs *reghdfe* to output results as our benchmark.

3.3. Results

Table 3a shows the OLS simulation results, which are the means and standard deviations over 100 simulations, of the following: the converged estimate of x, standard error, t statistics of null hypothesis that our estimate equal the true value (i.e., 2), number of iterations, difference between our estimate and that from *reghdfe*, and difference between the standard error reported from our algorithm and that from *reghdfe*. The models converge on average after 2.93 to 4.84 iterations depending on the specific data structure. Furthermore, when the unbalancedness of data increases, or the number of missing observations increases, it takes more iterations to converge. By changing the coefficients of time fixed effect in constructing explanatory variables and the range of time fixed effect in uniform distribution, we are able to test on the influence of remaining fixed effect and hence the convergence rate. As expected, when the dependence decreases, or the range of time fixed effect increases, the number of iteration increases. When the relative importance of time fixed effect changes, the final estimate and standard error do not change due to the elimination of remaining fixed effects in all variables. In Table 3b where the results of 2SLS models are shown, convergence requires fewer iterations when z is independent of time fixed effect and the opposite is true when the dependence of x on time fixed effect exclusively comes from z. The correlation of error terms in the two stages does not affect the results. If the clustered standard error is present, our algorithm obtains more precise estimates with fewer iterations. Finally, the last two columns of Table 3a and 3b show that our iterative results match well with the Stata default output, including standard errors that match with the Stata default outputs for all the variations of the model being examined here.

4. Empirical Example

4.1. Health Plan Type Effects on Health Care Utilization

We illustrate our algorithm using US employer-sponsored health insurance market data to extend the analyses in Ellis and Zhu (2016).

Ellis and Zhu (2016) estimate the health plan type effects on monthly health care treatment decisions. We use their data from the Truven Health Analytics MarketScan® Research Databases from 2007 to 2011 that contain detailed claims information for individuals insured by large employers in the US. The analysis sample contains 1.4 million individuals, ages 21-64, who are continuously insured from 2007 through 2011, with over 60 million treatment months for which they can assign an employer, a plan type, and a primary care physician (PCP). The extension in this paper is that we include 150,000 PCP fixed effects in addition to the 1.4 million individual and 3,000 county fixed effects, so that plan effects control not only for individual and geographic variation, but also in the specific PCPs seen by each consumer. Specifically, we estimate the following model.

$$Y_{it} = \mu PLAN_p + \beta X_{i,t-1} + \alpha_i + \delta_d + \gamma_c + \theta_t + \lambda_{EYF} + \varepsilon_{it} \qquad (3)$$

where Y_{it} is the indicator of doctor visit for consumer *i* in month *t*. The variables of interest, $PLAN_p$, are five plan dummies: EPO, HMO, POS, COMP, CDHP/HDHP. The omitted plan type is PPO, and hence the coefficients μ give the plan type effects as a difference from

PPOs.[6] Following Ellis and Zhu (2016), we instrument the endogenous plan type choice using the predicted plan type choice probabilities estimated from multinomial logit models at the household level.

In addition, we control for an enrollee's health status $X_{i,t-1,}$[7] enrollee fixed effects α_i, PCP fixed effects δ_d, employee county fixed effects γ_c, monthly time fixed effects θ_t, and employer*year*single/family coverage fixed effects λ_{EYF}. Finally, ε_{it} are error terms adjusted for clustering at the employer*year*single/family coverage level. Using equation (3), we identify the effects of plan innovations by the change in coverage for continuously eligible households.

4.2. Results

Due to the size of this data, it is impossible to estimate the model unless at least three dimensions of fixed effects are absorbed because apart from individual (about 1.4 million levels) and provider (about 150,000 levels) fixed effects, county fixed effects are also relatively high dimensional (about 3,000 levels). Also contributing to the challenge is the need to refine standard errors for 465 employer*year *coverage clusters.

Choosing how many fixed effects to absorb reflects a tradeoff between number of iterations and runtime. Generally, the more fixed effects absorbed, the more iterations needed for

[6] Plan type acronyms are: EPO (Exclusive Provider Organization), HMO (Health Maintenance Organization), POS (Point of Service, non-capitated), COMP (Comprehensive), and CDHP/HDHP (Consumer-Driven Health Plan/High-deductible Health Plan), and PPO (Preferred Provider Organization).
[7] We use prospective model risk score predicting total spending estimated from the prior twelve months of diagnoses to capture the patient's overall health status.

convergence, as convergence is generally harder to attain while the faster each iteration is by reducing the number of dummies variables in the model.

Table 4 shows that narrow network plans EPOs, HMOs and POS's reduce the probabilities of monthly provider contacts by 11.1%, 5.7%, 3.6%, respectively relative to PPO plans, while CDHP/HDHP plans are statistically insignificantly different (95% CI: -3.6% to 4.6%). Results suggest that narrow networks may be more effective than cost sharing in reducing health care utilization.

Figure 1 shows the convergence of estimates of five plan type effects and risk scores from regressing model (3) iteratively, with each iteration sequentially absorbing all five fixed effects. Number of iterations needed for convergence is significantly larger than that in the pseudo data, suggesting the important role of data structure in determining the speed of convergence. Examining the speed of convergence is a natural extension to this paper for future research.

5. Conclusions and Discussion

We present a new estimation algorithm that is particularly designed for models with multiple high-dimensional fixed effects and unbalanced panel. In essence, our algorithm absorbs fixed effects sequentially until they are asymptotically eliminated, which is straightforward and easy to implement. Monte Carlo simulations show that our approach matches results from estimation with fixed effect dummies in all the models. Furthermore, using our algorithm, it is feasible to estimate a model of health care utilization that involves 63 million observations from which we remove fixed effects for 1.4 million individuals,

150,000 distinct primary care doctors, 3,000 counties, 465 employer*year*single/family coverage dummies and 47 monthly time dummies.

In the Monte Carlo simulations, we also observe the changes of the estimate and standard error from the first iteration which provides insights for the remaining fixed effect bias and the pattern of convergence in our empirical case. For example, in OLS models, when the dependence of control variable on the time fixed effect decreases or the range of time fixed effect increases, the remaining fixed effect bias in the first iteration increases. On the other hand, in 2SLS models, if the instrument z is independent of time fixed effect, the remaining fixed effect bias is minimal and insignificant. In addition, if z depends on time fixed effect but such dependence is offset in x, where in extreme cases x is independent of time fixed effect, the remaining fixed effect bias increases and is extremely large in the extreme case. An interesting finding when we increase the noisiness of the models is that the remaining fixed effect in the first iteration is not affected, although the algorithm takes more iterations to converge and the converged estimates are less precise. This illustrates that the remaining fixed effect that cannot be eliminated after the first iteration dominates the idiosyncratic noise. Whether the standard error is clustered or not has no effect on estimates from the first iteration.

There is room to improve our algorithm. First, future studies could further investigate the convergence properties of our algorithm to improve its speed. Our simulation results offer some initial insights that convergence speed might mainly depend on relations between variables and fixed effects and the unbalanced structure of data. In addition, our algorithm might be modified to utilize prior information about the fixed effects to determine an

optimal order of sequential absorption that eliminates the fixed effects to the largest extent. Furthermore, certain numerical analysis techniques, e.g., Newton-Raphson Iteration, could be adopted to make our algorithm more efficient. Another extension of our paper would be to build analytical models to accommodate more than two high-dimensional fixed effects, allowing for a better understanding of the real data.

References:

Balázsi, László, László Mátyás, and Tom Wansbeek. (2015). "The Estimation of Multidimensional Fixed Effects Panel Data Models." *Econometric Reviews*. Available at: http://dx.doi.org/10.1080/07474938.2015.1032164

Cameron, Colin and Douglas L. Miller. (2015). "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, 50(2): 317-372.

Correia, Sergio. (2015). "REGHDFE: Stata Module for Linear and Instrumental-Variable/GMM Regression Absorbing Multiple Levels of Fixed Effects." Available at: https://ideas.repec.org/c/boc/bocode/s457874.html

Ellis, Randall P. and Wenjia Zhu. (2016). "Health Plan Type Variations in Spells of Health-Care Treatment." *American Journal of Health Economics*, 2(4): 399-430.

Guimaraes, Paulo, and Portugal, Pedro. (2010). "A Simple Feasible Alternative Procedure to Estimate Models with High-Dimensional Fixed Effects." *Stata Journal*, 10(4): 628-649.

Somaini, Paulo, and Frank A. Wolak. (2016). "An Algorithm to Estimate the Two-Way Fixed Effects Model." *Journal of Econometric Methods*, 5(1): 143-152.

Table 1. Comparison of Programs

		Clustered s.e.	*IV*	*One HDFE*	*2+ HDFE*	*Big Data*
Stata	*ivregress*	X	X			
	a2reg			X	X	
	reghdfe	X	X	X	X	
SAS	PROC GLM			X	X	X
	PROC SYSLIN		X			X
	PROC SURVEYREG	X				X
	TSLSFECLUS	X	X	X	X	X

Notes: Table summarizes the capability of various existing Stata (i.e., *ivregress*, *a2reg*, *reghdfe*) and SAS (i.e., PROC GLM, PROC SYSLIN, PROC SURVEYREG) commands, in comparison to our iterative algorithm (i.e., TSLSFECLUS), in handling models with the listed features. HDFE stands for high-dimensional fixed effect.

Table 2. Simulation Parameters for Two-way Fixed Effects Model

	M	ρ_{ZFEt}	ρ_{XFEt}	ρ_{uv}	$\bar{\theta}$	σ_u	Clustered s.e.
OLS	5000	0	80	0	100	10	N
	6000	0	80	0	100	10	N
	2000	0	80	0	100	10	N
	5000	0	**40**	0	100	10	N
	5000	0	**0**	0	100	10	N
	5000	0	**-80**	0	100	10	N
	5000	0	80	0	**10**	10	N
	5000	0	80	0	100	**100**	N
	5000	0	80	0	100	10	**Y**
2SLS	5000	40	40	0.6	100	10	N
	6000	40	40	0.6	100	10	N
	2000	40	40	0.6	100	10	N
	5000	**0**	**80**	0.6	100	10	N
	5000	**80**	**0**	0.6	100	10	N
	5000	**40**	**-40**	0.6	100	10	N
	5000	**60**	**-20**	0.6	100	10	N
	5000	**20**	**-20**	0.6	100	10	N
	5000	40	40	0.6	**10**	10	N
	5000	40	40	0.6	100	**100**	N
	5000	40	40	**0.2**	100	10	N
	5000	40	40	0.6	100	10	**Y**

Notes: Table shows the parameter inputs for simulating the two-way fixed effects model described in (1) in Section 2 of the text. Parameters are defined as: *M* (number of observations randomly selected to be dropped from the sample), ρ_{ZFEt} (dependence of the potential instrumental variable on time fixed effect), ρ_{XFEt} (dependence of the control variable on time fixed effect), ρ_{uv} (correlation between the endogenous variable and the error term), $\bar{\theta}$ (range of uniformly distributed time fixed effect), σ_u (standard deviation of error term in the dependent variable) and whether the clustered standard error (at the level of i) is present. Each row is a separate simulation for 100 times and the first row in each group of OLS/2SLS estimation shows the baseline parameter values. In each simulation, we change one or a pair of parameter values while fixing the others at their baseline values.

Table 3a. Simulation Results for OLS Models

Parameter	Coeff.	s.e.	t ($H_0: \hat{\beta} = 2$)	Iteration #	$\Delta_{\hat{\beta}} = \hat{\beta} - \hat{\beta}_0$	Δ_{se} = s.e. − s.e.$_0$
Baseline	1.9953	0.0471	-0.0963	4.05	-4.6E-05	-0.0002
	(0.0504)	(0.0022)	(1.0627)	(0.33)	(3.2E-05)	(0.0020)
$M = 6000$	2.0000	0.0509	0.0325	4.72	-5.6E-05	-0.0022
	(0.0537)	(0.0086)	(1.0261)	(0.75)	(0.0002)	(0.0085)
$M = 2000$	1.9989	0.0370	-0.0318	3.80	-5.4E-05	-3.0E-05
	(0.0382)	(0.0005)	(1.0296)	(0.42)	(3.0E-05)	(0.0003)
$\rho_{XFEt} = 40$	1.9954	0.0472	-0.0966	3.98	-4.1E-05	-9.5E-05
	(0.0504)	(0.0013)	(1.0628)	(0.20)	(3.6E-05)	(0.0012)
$\rho_{XFEt} = 0$	1.9954	0.0473	-0.0982	2.93	-4.4E-05	-9.8E-05
	(0.0504)	(0.0009)	(1.0618)	(0.26)	(2.6E-05)	(0.0005)
$\rho_{XFEt} = -80$	1.9954	0.0469	-0.1029	4.04	-3.4E-05	-0.0005
	(0.0504)	(0.0033)	(1.0632)	(0.40)	(7.4E-05)	(0.0032)
$\bar{\theta} = 10$	1.9954	0.0473	-0.0982	3.93	-4.4E-05	1.1E-05
	(0.0504)	(0.0009)	(1.0618)	(0.26)	(2.6E-05)	(0.0005)
$\sigma_u = 100$	1.9541	0.4734	-0.0973	4.84	-5.2E-05	0.0001
	(0.5038)	(0.0088)	(1.0618)	(0.39)	(2.9E-05)	(0.0054)
Clustered s.e.	1.9994	0.0054	-0.0181	3.96	-5.9E-05	-2.8E-05
	(0.0061)	(0.0011)	(1.5716)	(0.28)	(6.8E-05)	(0.0006)

Notes: Table shows the simulation results (including the coefficient and standard error estimates) on the 9 OLS models described in Table 2, using our iterative algorithm (i.e., TSLSFECLUS with *tol*=10^{-4}) and their comparison with results from *reghdfe*, $\hat{\beta}_0$ and s.e.$_0$. Results are means over 100 simulations with standard deviations shown in parentheses. The fixed parameters are $\{N, T, \rho_{ZFEi}, \rho_{XFEi}\}$={100, 100, 0, 0.2}. Baseline values for the rest of parameters are $\{M, \rho_{ZFEt}, \rho_{XFEt}, \rho_{uv}, \bar{\theta}, \sigma_u\}$= {5000, 0, 80, 0, 100, 10}. Clustered standard error correction, used in the last row, is at the level of *i*.

Table 3b. Simulation Results for 2SLS Models

Parameter	Coeff.	s.e.	t ($H_0: \hat{\beta} = 2$)	Iteration #	$\Delta_{\hat{\beta}} = \hat{\beta} - \hat{\beta}_0$	Δ_{se} = s.e. − s.e.$_0$
Baseline	1.9945	0.0501	-0.1040	4.03	-5.3E-05	-2.6E-05
	(0.0534)	(0.0009)	(1.0632)	(0.17)	(2.8E-05)	(0.0005)
$M = 6000$	1.9995	0.0558	0.0056	4.77	-5.1E-05	-0.0004
	(0.0571)	(0.0039)	(1.0154)	(0.49)	(2.9E-05)	(0.0038)
$M = 2000$	1.9988	0.0392	-0.0281	3.78	-4.8E-05	-2.5E-05
	(0.0392)	(0.0005)	(0.9971)	(0.41)	(2.8E-05)	(0.0003)
$\rho_{ZFEt} = 0$ $\rho_{XFEt} = 80$	1.9945	0.0501	-0.1040	2.94	-5.3E-05	-2.5E-05
	(0.0534)	(0.0009)	(1.0631)	(0.24)	(2.8E-05)	(0.0005)
$\rho_{ZFEt} = 80$ $\rho_{XFEt} = 0$	1.9945	0.0499	-0.1019	4.11	-4.8E-05	-0.0002
	(0.0534)	(0.0024)	(1.0641)	(0.40)	(6.0E-05)	(0.0021)
$\rho_{ZFEt} = 40$ $\rho_{XFEt} = -40$	1.9945	0.0501	-0.1040	4	-5.3E-05	-2.6E-05
	(0.0534)	(0.0009)	(1.0632)	(0)	(2.8E-05)	(0.0005)
$\rho_{ZFEt} = 60$ $\rho_{XFEt} = -20$	1.9945	0.0501	-0.1040	4	-5.3E-05	-2.6E-05
	(0.0534)	(0.0009)	(1.0632)	(0)	(2.8E-05)	(0.0005)
$\rho_{ZFEt} = 20$ $\rho_{XFEt} = -20$	1.9945	0.0501	-0.1039	4	-5.3E-05	-2.6E-05
	(0.0534)	(0.0009)	(1.0633)	(0)	(2.8E-05)	(0.0005)
$\bar{\theta} = 10$	1.9945	0.0501	-0.1040	3.82	-5.3E-05	-2.7E-05
	(0.0534)	(0.0009)	(1.0632)	(0.38)	(2.8E-05)	(0.0005)
$\sigma_u = 100$	1.9456	0.5009	-0.1030	4.73	-5.2E-05	-0.0003
	(0.5338)	(0.0092)	(1.0631)	(0.44)	(2.8E-05)	(0.0051)
$\rho_{uv} = 0.2$	1.9946	0.0501	-0.1062	4.03	-5.0E-05	-2.6E-05
	(0.0534)	(0.0009)	(1.0633)	(0.17)	(2.9E-05)	(0.0005)
Clustered s.e.	1.9991	0.0057	-0.1229	3.98	-4.9E-05	1.2E-05
	(0.0069)	(0.0010)	(1.3047)	(0.20)	(2.8E-05)	(0.0005)

Notes: Table shows the simulation results (including the coefficient and standard error estimates) on the 12 2SLS models described in Table 2, using our iterative algorithm (i.e., TSLSFECLUS with *tol*=10^{-4}) and their comparison with results from *reghdfe*, $\hat{\beta}_0$ and s.e.$_0$. Results are means over 100 simulations with standard deviations shown in parentheses. The fixed parameters are {*N*, *T*, ρ_{ZFEi}, ρ_{XFEi}}={100, 100, 0, 0.2}. Baseline values for the rest of parameters are{$M, \rho_{ZFEt}, \rho_{XFEt}, \rho_{uv}, \bar{\theta}, \sigma_u$}= {5000, 40, 40, 0.6, 100, 10}. Clustered standard error correction, used in the last row, is at the level of *i*.

Table 4. Health Plan Type Effects on Health Care Utilization

	Pr(any visit)	
EPO	-0.111	**
	(0.053)	
HMO	-0.057	***
	(0.015)	
POS	-0.036	***
	(0.009)	
COMP	0.051	
	(0.043)	
CDHP/HDHP	0.005	
	(0.021)	
Prospective risk score	0.024	***
	(0.001)	
Dep. Var. Mean	0.321	
Observations	62,899,584	

Notes: Table shows the 2SLS estimates of health plan type effects on the probability of seeking care. Plan acronyms are defined as: EPO (Exclusive Provider Organization), HMO (Health Maintenance Organization), POS (Point of Service, non-capitated), COMP (Comprehensive), and CDHP/HDHP (Consumer-Driven Health Plan/High-deductible Health Plan). PPO (Preferred Provider Organization) is the omitted plan type. The annual prospective risk score is the predicted total spending using the prior 12 months of diagnoses. Regression also controls for individual fixed effects, PCP fixed effects, employee county fixed effects, employer*year*family coverage fixed effects, and monthly time fixed effects. Standard errors are adjusted for clustering at the level of employer-year-single/family coverage type. *** = $p<0.01$, ** = $p<0.05$, * = $p<0.10$.

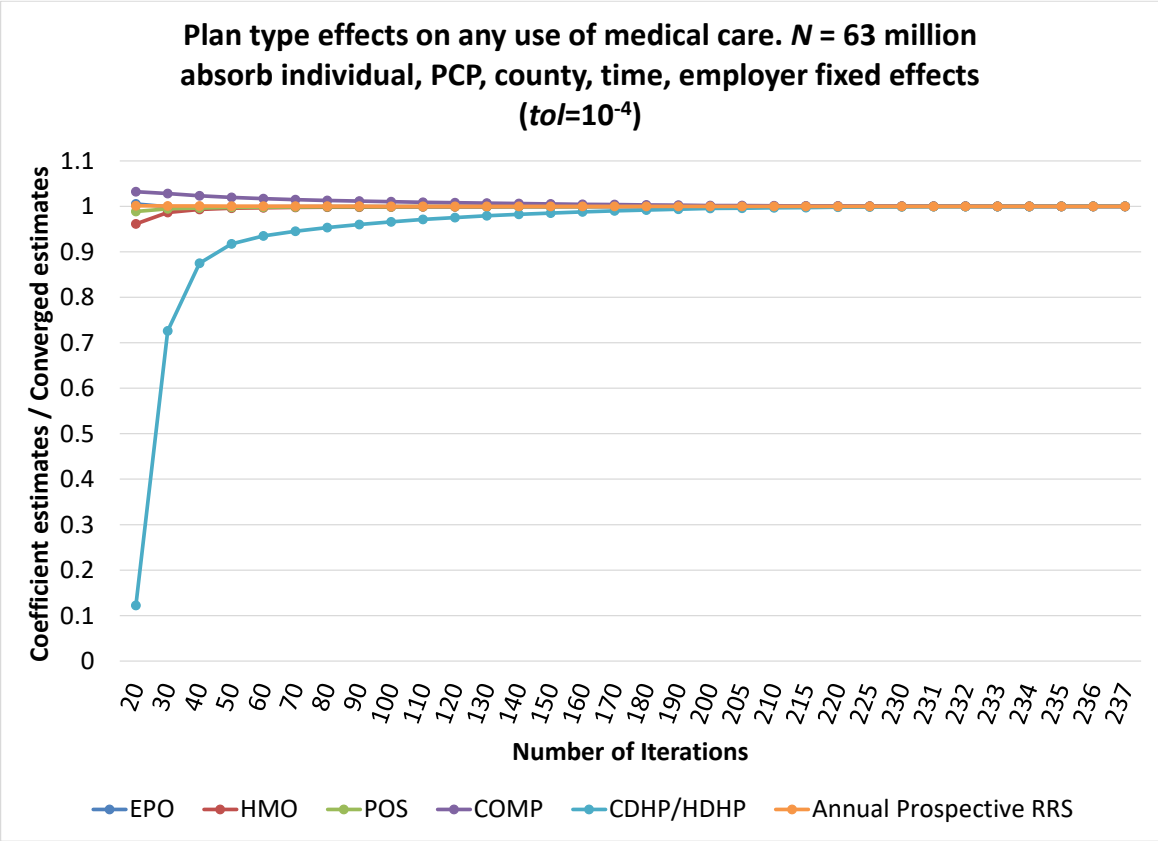

Figure 1. Convergence of Estimates of Health Plan Type Effects on Health Care Utilization

Notes: Figure shows the convergence of 2SLS estimates of health plan type effects on the probability of seeking care. Each plan type estimate is normalized by its value at the last iteration (i.e., iteration 237). PPO (Preferred Provider Organization) is the omitted plan type. The underlying regression controls for individual fixed effects, PCP fixed effects, employee county fixed effects, employer*year*family coverage fixed effects, and monthly time fixed effects, a prospective model risk score predicting total spending using the prior 12 months of diagnoses.

Appendix A: Proof of Theorems and Corollary

Proof of Theorem 1:

Define the T-by-1 vector of remaining time fixed effects at k-th iteration as:

$$\theta^{(k)} = \begin{bmatrix} \theta_1^{(k)} \\ \vdots \\ \theta_T^{(k)} \end{bmatrix}$$

$\forall r = 1,2, \dots, T$, from equation (2)

$$\begin{aligned} \theta_r^{(k+1)} &\equiv \frac{1}{\|N_r\|} \sum_{i \in N_r} \frac{1}{\|T_i\|} \sum_{t \in T_i} \theta_t^{(k)} \\ &= \frac{1}{\sum_{j=1}^{N} 1(j \in N_r)} \sum_{i=1}^{N} 1(i \in N_r) \cdot \left[\frac{1}{\sum_{s=1}^{T} 1(s \in T_i)} \sum_{t=1}^{T} 1(t \in T_i) \theta_t^{(k)} \right] \\ &= \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{1(i \in N_r) 1(t \in T_i)}{\sum_{j=1}^{N} 1(j \in N_r) \sum_{s=1}^{T} 1(s \in T_i)} \theta_t^{(k)} \\ &= \sum_{t=1}^{T} \lambda_{rt} \theta_t^{(k)} \end{aligned}$$

Then the linear system of the remaining fixed effect between adjacent iterations is as follow:

$$\theta^{(k+1)} = \Lambda \theta^{(k)}$$

where the linear transformation T-by-T matrix Λ with (r,t) entry:

$$\lambda_{rt} \equiv \sum_{i=1}^{N} \frac{1(i \in N_r) 1(t \in T_i)}{\sum_{j=1}^{N} 1(j \in N_r) \sum_{s=1}^{T} 1(s \in T_i)}$$

In other word, the fixed effects at (k+1)th iteration are the weighted averages of fixed effects from k-th iteration and the weights satisfy the following two conditions.

1. Summations within rows are 1, $\forall r$

$$\sum_{t=1}^{T}\lambda_{rt} = \sum_{t=1}^{T}\sum_{i=1}^{N}\frac{1(i\in N_r)1(t\in T_i)}{\sum_{j=1}^{N}1(j\in N_r)\sum_{s=1}^{T}1(s\in T_i)}$$
$$= \sum_{i=1}^{N}\frac{1(i\in N_r)\sum_{t=1}^{T}1(t\in T_i)}{\sum_{j=1}^{N}1(j\in N_r)\sum_{s=1}^{T}1(s\in T_i)}$$
$$= \sum_{i=1}^{N}\frac{1(i\in N_r)}{\sum_{j=1}^{N}1(j\in N_r)} = 1$$

2. $0 \le \lambda_{rt} \le 1, \forall r,t$. Notice that under Assumption 1, this condition converts to $0 < \lambda_{rt} \le 1, \forall r,t$. To see that $\lambda_{rt} \neq 0$, use the relation $1(t\in T_i) = 1(i\in N_t)$.

$$\lambda_{rt} = \sum_{i=1}^{N}\frac{1(i\in N_r)1(i\in N_t)}{\sum_{j=1}^{N}1(j\in N_r)\sum_{s=1}^{T}1(s\in T_i)} \neq 0$$

By assuming T finite, define

$$M_k = \max\{\theta_1^{(k)},\dots,\theta_T^{(k)}\}$$
$$m_k = \min\{\theta_1^{(k)},\dots,\theta_T^{(k)}\} \quad \forall k$$

Then there exists $\bar{r}$, a function of k, and $\bar{r}\in\{1,2,\dots,T\}$ s.t.

$$M_{k+1} = \theta_{\bar{r}}^{(k+1)} = \sum_{t=1}^{T}\lambda_{\bar{r}t}\theta_t^{(k)}$$
$$\le \sum_{t=1}^{T}\lambda_{\bar{r}t}M_k$$
$$= M_k$$

Inequality holds by condition 2 and the last equality results from condition 1.

Similarly we have $m_{k+1} \ge m_k$

$$\Rightarrow M_0 \ge M_1 \ge \cdots \ge M_k \ge M_{k+1} \ge m_{k+1} \ge m_k \ge \cdots \ge m_1 \ge m_0$$

So $\{M_k\}_{k=1}^{\infty}$ and $\{m_k\}_{k=1}^{\infty}$ are both monotonic and bounded. By Monotone Convergence Theorem, their limits exist and are finite.

$$\lim_{k\to\infty} M_k = M \text{ and } \lim_{k\to\infty} m_k = m$$

WTS: M = m, so fixed effects converge to constant.

Proof:

By $\lim_{k\to\infty} M_k = M$ and $\lim_{k\to\infty} m_k = m$,

$$\forall \varepsilon, \quad \exists L_1 \ s.t. \forall k > L_1, \ |M_k - M| < \frac{\varepsilon}{3}$$

$$\exists L_2 \ s.t. \forall k > L_2, \ |m_k - m| < \frac{\varepsilon}{3}$$

$$\exists L_3 \ s.t. \forall k_1, k_2 > L_3, \ \left|M_{k_1} - M_{k_2}\right| < \frac{\varepsilon}{6}\lambda$$

$$\exists L_4 \ s.t. \forall k_1, k_2 > L_4, \ \left|m_{k_1} - m_{k_2}\right| < \frac{\varepsilon}{6}\lambda$$

where $\lambda \equiv \min\{\lambda_{rt}: \lambda_{rt} \neq 0\}$. By T being finite, $\lambda > 0$. Note that under Assumption 1, $\lambda = \min\{\lambda_{rt}\} > 0$.

Let $L = \max\{L_1, L_2, L_3, L_4\} + 1$, then $\forall k_1, k_2 \geq L$

$$\begin{cases} \left|M_{k_1} - M\right| < \frac{\varepsilon}{3} \\ \left|m_{k_1} - m\right| < \frac{\varepsilon}{3} \\ \left|M_{k_1} - M_{k_2}\right| < \frac{\varepsilon}{6}\lambda \\ \left|m_{k_1} - m_{k_2}\right| < \frac{\varepsilon}{6}\lambda \end{cases}$$

Without loss of generality, take $k_1 = L, k_2 = L + 1$. If at L-th iteration, time fixed effects are constant, it is trivial that this linear system is stabilized at this specific fixed point. Otherwise, there are two cases for $\theta^{(L)}$.

(1) $\exists t_0 \ s.t. m_L < \theta_{t_0}^{(L)} < M_L$

(2) $\theta_t^{(k)} = \begin{cases} M_k \\ m_k \end{cases} \forall t$

In case (1), suppose $\bar{r}_L$ and $\underline{r}_L$ are such that $M_{L+1} = \sum_{t=1}^{T} \lambda_{\bar{r}_L t}\theta_t^{(L)}$ and $m_{L+1} = \sum_{t=1}^{T} \lambda_{\underline{r}_L t}\theta_t^{(L)}$.

$$
\begin{aligned}
\left|M_{k_1} - M_{k_2}\right| &= M_L - M_{L+1} \\
&= M_L - \sum_{t=1}^{T} \lambda_{\bar{r}_L t} \theta_t^{(L)} \\
&\geq M_L - \sum_{t \neq t_0} \lambda_{\bar{r}_L t} M_L - \lambda_{\bar{r}_L t_0} \theta_{t_0}^{(L)} \\
&= \lambda_{\bar{r}_L t_0} (M_L - \theta_{t_0}^{(L)}) \\
&\geq \lambda (M_L - \theta_{t_0}^{(L)}) \\
&\Rightarrow M_L - \theta_{t_0}^{(L)} < \frac{\varepsilon}{6}
\end{aligned}
$$

$$
\begin{aligned}
\left|m_{k_1} - m_{k_2}\right| &= m_{L+1} - m_L \\
&= \sum_{t=1}^{T} \lambda_{\underline{r}_L t} \theta_t^{(L)} - m_L \\
&\geq \sum_{t \neq t_0} \lambda_{\underline{r}_L t} m_L + \lambda_{\underline{r}_L t_0} \theta_{t_0}^{(L)} - m_L \\
&= \lambda_{\underline{r}_L t_0} (\theta_{t_0}^{(L)} - m_L) \\
&\geq \lambda (\theta_{t_0}^{(L)} - m_L) \\
&\Rightarrow \theta_{t_0}^{(L)} - m_L < \frac{\varepsilon}{6}
\end{aligned}
$$

By these two inequalities: $M_L - m_L = M_L - \theta_{t_0}^{(L)} + \theta_{t_0}^{(L)} - m_L < \frac{\varepsilon}{3}$

In case (2), simply let t_0 be such that $\theta_{t_0}^{(L)} = m_L$ in either one of the inequalities, then $M_L - m_L < \frac{\varepsilon}{6} < \frac{\varepsilon}{3}$.

$$
\begin{aligned}
|M - m| &\leq |M - M_L| + |M_L - m_L| + |m_L - m| \\
&< \frac{\varepsilon}{3} + \frac{\varepsilon}{3} + \frac{\varepsilon}{3} \\
&= \varepsilon
\end{aligned}
$$

So M = m. ■

Proof of Theorem 2:

Model specification in matrix form:

$$Y = X\beta + D_\alpha \alpha + D_\theta \theta + u$$

The estimator of β from this regression is the LSDV estimator, $\hat{\beta}_{LSDV}$.

In 1st iteration of our algorithm:

1) Demean over i, the index of fixed effect α

$$\begin{aligned} M_\alpha Y &= M_\alpha X\beta + M_\alpha D_\theta \theta + M_\alpha u \\ &= M_\alpha X\beta + D_\theta \theta - P_\alpha D_\theta \theta + M_\alpha u \\ &= M_\alpha X\beta + D_\theta \theta + D_\alpha \tilde{\alpha} + M_\alpha u \end{aligned}$$

where *M.* denotes the annihilator matrix projecting the variables to the orthogonal space of some corresponding dummy variables, e.g. $M_\alpha = I - P_\alpha = I - D_\alpha (D_\alpha' D_\alpha)^{-1} D_\alpha'$. By Frisch-Waugh-Lovell (FWL) Theorem, the estimator from this regression model is identical to $\hat{\beta}_{LSDV}$.

2) Demean over t, the index of fixed effect θ

$$\begin{aligned} M_\theta M_\alpha Y &= M_\theta M_\alpha X\beta + M_\theta D_\alpha \tilde{\alpha} + M_\theta M_\alpha u \\ &= M_\theta M_\alpha X\beta + M_\theta M_\alpha D_\theta \theta + M_\theta M_\alpha u \end{aligned}$$

$M_\theta M_\alpha D_\theta \theta = D_\alpha \tilde{\alpha} + D_\theta \tilde{\theta}$ represents the remaining fixed effects after 1 iteration.

$$\tilde{\alpha} = -(D_\alpha' D_\alpha)^{-1} D_\alpha' D_\theta \theta$$

$$\tilde{\theta} = (D_\theta' D_\theta)^{-1} D_\theta' P_\alpha D_\theta \theta$$

The above equations are consistent with the fixed effect updating formula (2). Applying FWL Theorem once again on the above model, the estimator of β remains identical to $\hat{\beta}_{LSDV}$. Due to the remaining two-way fixed effects, we can apply FWL Theorem continuously for each demeaning in our iteration process. The estimator from each iteration with the correct specification of regression model should be the same as $\hat{\beta}_{LSDV}$.

In k-th iteration of our algorithm:

The correct specification of model is

$$\tilde{Y}^{(k)} = \tilde{X}^{(k)}\beta + \tilde{D}_{\theta}^{(k)}\theta + \tilde{u}^{(k)}$$

where

$$\tilde{X}^{(k)} = (M_{\theta}M_{\alpha})^{k}X$$

$$\tilde{Y}^{(k)} = (M_{\theta}M_{\alpha})^{k}Y$$

$$\tilde{D}_{\theta}^{(k)} = (M_{\theta}M_{\alpha})^{k}D_{\theta}$$

$$\tilde{u}^{(k)} = (M_{\theta}M_{\alpha})^{k}u$$

Denote $\widetilde{M}^{(k)} = I - \tilde{D}_{\theta}^{(k)}(\tilde{D}_{\theta}^{(k)\prime}\tilde{D}_{\theta}^{(k)})^{-1}\tilde{D}_{\theta}^{(k)\prime}$ and transform the model to

$$\widetilde{M}^{(k)}\tilde{Y}^{(k)} = \widetilde{M}^{(k)}\tilde{X}^{(k)}\beta + \widetilde{M}^{(k)}\tilde{u}^{(k)}$$

and by FWL Theorem, the estimator with absorption of remaining fixed effects is $\hat{\beta}_{RFE}^{(k)} = \left(\tilde{X}^{(k)\prime}\widetilde{M}^{(k)}\tilde{X}^{(k)}\right)^{-1}\tilde{X}^{(k)\prime}\widetilde{M}^{(k)}\tilde{Y}^{(k)} \equiv \hat{\beta}_{LSDV},\ \forall$k.

However, under our regression specification without remaining fixed effects, the OLS estimator from this iteration is as following:

$$\hat{\beta}^{(k)} = \left(\tilde{X}^{(k)\prime}\tilde{X}^{(k)}\right)^{-1}\tilde{X}^{(k)\prime}\tilde{Y}^{(k)}$$

We only need to show $\hat{\beta}^{(k)} \to \hat{\beta}_{RFE}^{(k)} = \hat{\beta}_{LSDV}$, as k $\to \infty$.

Proof:

The remaining fixed effects $\tilde{D}_{\theta}^{(k)}\theta = (M_{\theta}M_{\alpha})^{k}D_{\theta}\theta = D_{\alpha}\tilde{\alpha}^{(k)} + D_{\theta}\tilde{\theta}^{(k)} \to 0$ for any arbitrary θ, as a result of Theorem 1. It implies that $\tilde{D}_{\theta}^{(k)} \to 0$ as k $\to \infty$. Hence $\widetilde{M}^{(k)} \to$ I and $\hat{\beta}^{(k)} \to \hat{\beta}_{RFE}^{(k)}$. ■

Proof of Corollary 1:

When the estimator from our algorithm converges to the LSDV estimator, the regression model

$$\tilde{y}^{(\infty)} = \tilde{x}^{(\infty)}\beta + \tilde{u}^{(\infty)}$$

is correctly specified and the estimator $\hat{\beta}^{(\infty)}$ is unbiased. The bias of remaining fixed effects is 0, which indicates that the remaining fixed effects are eliminated or equivalently there is no longer omitted fixed effect variables in the regression error. ■

Appendix B: Implementation of TSLSFECLUS Algorithm

B.1. Implementation Steps

Our algorithm is programmed in SAS for ease of implementation. The main macro that performs the iterative procedure is TSLSFECLUS. This macro can accommodate a wide range of model features such as endogeneity, cluster standard error correction, and multiple high-dimensional fixed effects. In addition, it allows multiple specifications that differ only in their dependent variables to be estimated in a single call. Finally, the macro automatically outputs the number of iterations needed for model convergence together with the model estimates. The macro mainly contains the following four steps.

1): Given model specification, identify multiple high-dimensional fixed effects to absorb. Set the values of maximum number of iteration $Maxiter$ and the tolerance level tol.

2): Absorb fixed effects from all dependent and explanatory (including instrumental) variables, one by one until all the fixed effects are absorbed once. Save standardized data S_1, and the estimated parameters of interest from model S_1, labeled as $\{\hat{\beta}^{(1)}{}_k\}_{k=1,2,\dots,K}$.

3): Repeat step 2) and record S_2, and obtain $\{\hat{\beta}^{(2)}{}_k\}_{k=1,2,\dots,K}$ from estimating the model using S_2.

4): Calculate $|\Delta_2| = \max\{|\frac{\hat{\beta}^{(2)}{}_k - \hat{\beta}^{(1)}{}_k}{\hat{\beta}^{(1)}{}_k}|\}_{k=1,2,\dots,K}$, the maximum absolute value of percentage difference between adjacent iterations among *K* estimated parameters of interest. If $|\Delta_2| < tol$, then stop here and report coefficient estimates $\{\hat{\beta}^{(2)}{}_k\}_{k=1,2,\dots,K}$; otherwise, repeat step 2) until $|\Delta_i| = \max\{|\frac{\hat{\beta}^{(i)}{}_k - \hat{\beta}^{(i-1)}{}_k}{\hat{\beta}^{(i-1)}{}_k}|\}_{k=1,2,\dots,K} < tol$ or the maximum number of iterations have been reached. The reported number of iteration = $\min\{i = \{i||\Delta_i| < tol\}, Maxiter\}$.

B.2. Sample Call of TSLSFECLUS

Below is a sample call of our two macros in SAS. The second macro can be called directly if only one iteration is desired, such as if there is only one high-dimensional fixed effect:

Libname junk “directory for storing temporary data sets”;

```
%auto_iter(
indsn=in_data,          /* input data */
tol=0.0001,             /* tolerance level for convergence */
maxiter=100,            /* maximum number of iteration */
```

```
betasefinal=out_beta,       /* output data for storing estimates from all iterations */
fevarcount=2,               /* number of absorbed fixed effects */
tempdir=junk                /* directory for storing temporary data sets */
);

*which calls the following core macro iteratively;
%TSLSCLUS_iterFE(
runtitle='two-way FE model',    /* running title */
indata = &indsn.,               /* input data for each iteration: &indsn. for the first
                                   iteration, standardized data for subsequent iterations */
depvar = outcome,               /* dependent variable */
endog = y,                      /* endogenous variable */
inst = z,                       /* instrumental variable */
exog = x,                       /* exogenous variable */
fe = i c t,                     /* variables defining absorbed fixed effects */
FE_iter = 1,                    /* incremental on iteration number */
cluster = c,                    /* variable defining cluster level */
othervar = w,                   /* other variables to be carried along to final dataset for
                                   final analysis*/
tempdir = junk,                 /* directory for storing temporary data sets */
regtype = TSLS,                 /* TSLS or OLS */
showmeans = no,                 /* yes or no to showing sample summary statistics */
showrf = no,                    /* yes or no to showing reduced form results of TSLS
                                   model */
showols = no,                   /* yes or no to OLS without cluster correction */
dosurveyreg = no,               /* yes or no to doing PROC SURVEYREG */
wide = no,                      /* yes or no to wide format output table */
estresult=betasecurr            /* data set for outputting estimates from current iteration
                                   */
);
```

B.3. Clustering Standard Errors

In many economic settings, standard errors are not necessarily independent but correlated within groups (e.g., schools, households, etc.), a phenomenon known as "clustered standard errors". For example, student performance may be correlated within schools, and health spending is likely to be correlated within households. Suppose we allow errors to cluster at *G* level. Let *g* denote g^{th} element in *G*. Following Cameron and Miller (2015), clustered errors can be expressed as:

$$E\left(u_{itg}u_{jsg'}|x_{itg}, x_{jsg'}\right) = 0 \text{ unless } g = g'$$

Then the cluster-robust variance estimator (CRVE) can be written as:

$$CRVE = (\tilde{X}'\tilde{X})^{-1} \sum_{g=1}^{G} v_g' v_g \, (\tilde{X}'\tilde{X})^{-1}$$

where $\tilde{X}$ is a $(\sum_i T_i) \times K$ matrix of the demeaned Xs of the converged model, $v_g = \sum_{it \in g} e_{it} \tilde{x}_{it}$ and $e_{it} = \tilde{y}_{it} - \hat{\beta}' \tilde{x}_{it}$.

In the empirical estimation, we calculate cluster-robust standard errors by applying the above formula to the converged model where $(\tilde{x}_{it}, \tilde{y}_{it})$ are the demeaned values.