Practical issues in the estimation of high-dimensional fixed effects models

Miguel Portela Paulo Guimarães

Porto, June 26, 2022

Introduction

- we live on the era of “big data”
- availability of microdata for researchers is better than ever
- easy to gain access to large administrative data sets
- these ”large data sets” open up research possibilities
- but they also pose many technical challenges

Introduction

- large and more detailed data sets make it possible to control for different sources of unobserved heterogeneity
- examples are:
  - Employer-employee level data
  - Hospital-doctor-patient data
  - School-class-teacher-student level data
  - Loan-borrower-bank data
- these data are sometimes described as multilevel or hierarchical
- repeated observations allow for the introduction of additional error components that account for unobserved heterogeneity

Introduction

- if error components are uncorrelated with observed explanatory variables we can use mixed models (error components are treated as random effects)
- if error components are correlated with observed explanatory variables then they need to be treated as fixed effects
- introducing fixed effects in a linear regression amounts to modelling the error component using dummy variables
- with only one error component this is the usual fixed effects panel data regression

Single fixed effects model

- suppose that you want to estimate the model

$$y_{it} = \mu + \mathbf{x}'_{it}\beta + \mathbf{u}_i\eta + \alpha_i + u_{it}$$

where you have observations of multiple individuals observed over time. The subscript i indexes individual and t stands for time. Strict exogeneity, $E(u_{it}|\mathbf{x}'_{it}, \mathbf{u}_i, \alpha_i) = 0$, is assumed.

$\mathbf{x}_{it}$ - time-varying individual level observed explanatory variables

$\mathbf{u}_i$ - time invariant individual level observed explanatory variables

α_i - time invariant individual level unobserved explanatory variables

Single fixed effects model

- the vector β can be estimated by running the regression

$$y_{it} - \overline{y}_i = (\mathbf{x}'_{it} - \overline{\mathbf{x}}'_i)\beta + \widetilde{u}_{it}$$

- the estimates for β are the same as if we had included a dummy variable per individual
- time invariant individual level observed variables are absorbed
- this will work regardless of the number of individuals (high-dimensional)

Two high dimensional fixed effects

- an example: the AKM wage regression:

$$y_{it} = \mu + \mathbf{x}'_{it}\beta + \mathbf{w}_{j(i,t)}\gamma + \mathbf{u}_i\eta + \mathbf{q}_{j(i,t)}\rho + \alpha_i + \phi_{j(i,t)} + \mu_t + u_{it}$$

y_{it} - wage of worker i at time t

$\mathbf{x}_{it}$ - time-varying worker observed explanatory variables

$\mathbf{w}_{j(i,t)}$ - time-varying firm observed explanatory variables

$\mathbf{u}_i$ - time invariant worker observed explanatory variables

$\mathbf{q}_{j(i,t)}$ - time invariant firm observed explanatory variables

α_i - time invariant worker unobserved explanatory variables

$\phi_{j(i,t)}$ - time invariant firm unobserved explanatory variables

μ_t - unobserved time effect

u_{it} - usual error term

Two high dimensional fixed effects

- in practice this is what is estimated:

$$y_{it} = \mu + \mathbf{x}'_{it}\beta + \mathbf{w}_{j(i,t)}\gamma + \theta_i + \psi_{j(i,t)} + \mu_t + u_{it}$$

where $\theta_i \equiv \alpha_i + \mathbf{u}_i\eta$ and $\psi_j \equiv \mathbf{q}_j\rho + \phi_j$

Estimation with two FEs

- rewrite the two fixed-effects model in matrix form:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}_1\alpha + \mathbf{D}_2\gamma + \epsilon$$

- $\mathbf{D}_1$ is $n \times G_1$ and $\mathbf{D}_2$ is $n \times G_2$ and both G_1 and G_2 are large numbers
- direct estimation of this model is complicated

Estimation with two FEs

- but a "zigzag" approach is simple to implement:

$$\left[\begin{array}{l} \beta^{(j+1)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y} - \mathbf{D}_1\alpha^{(j)} - \mathbf{D}_2\gamma^{(j)}\right) \\ \alpha^{(j)} = \left(\mathbf{D}_1'\mathbf{D}_1\right)^{-1}\mathbf{D}_1'\left(\mathbf{Y} - \mathbf{X}\beta^{(j)} - \mathbf{D}_2\gamma^{(j)}\right) \\ \gamma^{(j)} = \left(\mathbf{D}_2'\mathbf{D}_2\right)^{-1}\mathbf{D}_2'\left(\mathbf{Y} - \mathbf{X}\beta^{(j)} - \mathbf{D}_1\alpha^{(j)}\right) \end{array}\right]$$

- take advantage of the Frisch-Waugh-Lovell theorem
- standard errors (clustered or not) can also be easily calculated
- the degrees of freedom of the regression are

$$dof = n - (k + G_1 + G_2 - M)$$

where M is the number of coefficients dropped (or mobility groups)

Extending the 2hdfe linear model

- this iterative estimation technique can be extended to several sets of fixed effects

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}\alpha + \epsilon$$

where $\mathbf{D} = [\mathbf{D}_1\mathbf{D}_2...\mathbf{D}_F]$ is a matrix containing several fixed effects or interacted fixed effects. Eg:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}\alpha + \mathbf{LD}\gamma + \epsilon$$

LD are interactions between the fixed effect and the variables

- we can use the same approach as for 2hdfe to estimate these models
- with more than 2 hdfes calculation of the d.o.f. is complicated. But do we really care?

General considerations

- remember that with short-panels the estimates of the fes may be inconsistent
- but averages, kernel-densities, etc should be fine!
- with 2 hdfe estimates of the fes may only be compared **within** each mobility group. That is why researchers work with the "largest connected set"
- with more than 2 fes one can find a connected set using the algorithm of Weeks and Williams (1964). Within this set estimates of fes are comparable
- correlation between the estimates of the fixed effects may be biased (eg: limited mobility)
- is mobility endogenous?

What about nonlinear models?

- many nonlinear models are estimated using iterative algorithms based on linear regression
- an example are Generalized Linear Models estimated by Iteratively Reweighted Least Squares (IRLS)
- another example are nonlinear models that can be estimated recursively using linear regression (the NLS algorithm)
- ability to estimate does not translate into consistency of estimators

GLM models can be estimated by IRLS

- examples of GLM models where we can add fes:
  - Poisson regression
  - logit regression
  - probit regression
  - cloglog regression
  - negative binomial
  - gamma regression
- we need to worry about the incidental parameter problem (IPP)
- the IPP is not a concern for Poisson regression
- There are approaches to correct the IPP bias

Advice for estimation

- prepare a "clean and lean dataset"
  - Keep only needed variables
  - Drop missing values
  - Drop singletons
- if you plan on doing secondary analysis of fes restrict your data to a "connected set"
- if you use clustered standard errors make sure the number of clusters is high enough (+50)
- make sure you understand well the package you are using

Commands for estimation of high dimensional models

Stata

- `reghdfe` - OLS with hdfe - by Sergio Correia
- `ppmlhdfe` - PPML with hdfe - by Sergio Correia, Paulo Guimarães and Thomas Zylkin

R

- `fixest` - OLS and GLM with hdfe - by Laurent Bergé

Julia

- `FixedEffectModels` - OLS with hdfe - by Matthieu Gomez

Python

- `pyhdfe` - OLS with hdfe - by Jeff Gortmaker

References

- Abowd, J, Kramarz, F. and Margolis, D. "High wage workers and high wage firms", Econometrica, vol. 67(2), 251-233 (1999).
- Guimarães, P. and Portugal, P. "A simple feasible procedure to fit models with high-dimensional fixed effects", the Stata Journal, 10(4) 628-649, 2010.
- Correia, S. "Singletons, cluster robust standard errors and fixed effects: a bad mix", unpublished paper.