Using Stata to estimate nonlinear models with high-dimensional fixed effects

Paulo Guimaraes[1,2]

[1] Banco de Portugal
[2] Universidade do Porto

Portuguese Stata UGM - Sept 15, 2017

"More and more data"?

- availability of microdata for researchers is increasing fast
- easy to gain access to very large data sets
- these "large data sets" open up research possibilities
- they also pose many technical challenges
- an important limitation is the lack of tools to efficiently explore large data sets

The response

- Stata made significant improvements to respond to the need to work with larger data sets
  - introduction of Mata
  - Stata MP
  - increase in Stata limits
  - faster code for many ados
  - plugins
- and the Stata community also offered contributions
  - `parallel` - by George Vega Yon
  - `ftools` - by Sergio Correia
  - `gtools` - by Mauricio Caceres Bravo

What about regressions for high-dimensional data?

- Stata has significantly expanded methods for panel/longitudinal data
- but it still lacks command for dealing with regressions with multiple fixed effects
- many user-written packages for linear regression:
  - `areg` by Amine Ouazad
  - `reg2hdfe` by Paulo Guimaraes
  - `gpreg` by Johannes F. Schmieder
  - `felsdvreg` by Thomas Cornelissen
  - `reghdfe` by Sergio Correia
- `reghdfe` is the gold standard!
- it is very fast, allows weighs, and it handles multiple fixed effects and interactions

What about nonlinear regression models with multiple fixed effects?

- There are theoretical challenges
  - are the relevant parameters identifiable?
  - does the solution exist?
  - is the incidental parameter problem "biting"?
- and there are technical challenges ...
  - what algorithms to use?
  - are the approaches computationally feasible?
  - are algorithms fast enough for large data sets?

But there is hope for many nonlinear models

- `reghdfe` does a great job for linear regression
- makes possible estimation of nonlinear models by iterative algorithms based on linear regression
- a good example are Generalized Linear Models - can be efficiently estimated by Iteratively Reweighted Least Squares
- another example are nonlinear models that can be estimated recursively using linear regressions

GLM - Generalized Linear Models

- GLM models can be estimade by IRLS as

$$\left(\mathbf{X}'\mathbf{W}^{(r-1)}\mathbf{X}\right)^{-1} = \mathbf{X}'\mathbf{W}^{(r-1)}\mathbf{z}^{(r-1)}$$

- Examples of GLM models are:
  - Poisson regression
  - logit regression
  - probit regression
  - cloglog regression
  - negative binomial
  - gamma
  - All of these (and more) can be estimated by IRLS
  - It is a simple matter to add hdfes!
  - `poi2hdfe` is an example for Poisson with 2 hdfes

Some examples

- Example 1 - Poisson regression with 2 hdfes
- Example 2 - cloglog with 2 hdfes

Regression with peer effects

- a regression with peer effects (Arcidiacono et al, 2012) can be written as

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

- the regression is non linear
- estimation can be implemented by alternating between estimation of β, γ and estimation of α
- conditional on α the problem becomes linear
- easy to add other fixed effects

An example of peer regression

- Example - regression with peer effects

Conclusion

- it is possible to add fixed effects to some nonlinear models
- Poisson regression is probably the easiest application
- but we should worry about existence of a solution
- ability to estimate does not translate into consistency of estimators
- should understand better how long a panel needs to be
- estimation on large data sets likely to be a slow process