Estimation of High-Dimensional Models

Paulo Guimarães

Outline

Introduction

LRM - 1FE

LRM - 2FE

LRM - 3FE

Degrees of Freedom

Nonlinear Models

Estimation of High-Dimensional Models

Paulo Guimarães

Portuguese Stata Users Group
University of Minho, Braga 2010

Outline

- Introduction
  - What do I mean by "high-dimensional models"?
  - How can we estimate "high-dimensional models"?
- The Linear Regression Model
  - 1 fixed effect
  - 2 fixed effects
  - 3 fixed effects
- Identification of the fixed effects
- Non-linear regression models with fixed effects

Introduction

- Estimation of models with many observations and variables poses new challenges.
- Conventional estimation methods will not work.
- A case in point is estimation of models with high-dimensional fixed effects.
- With high-dimensional models explicit introduction of dummy variables to account for fixed effects is not an option.
- With one fixed effect there are other solutions:
  - Condition out the fixed effects (eg: linear regression, poisson, logistic regression)
  - use a modified iterative algorithm for maximization (see Greene(2004))

Introduction

What happens if:

- We can not condition out the fixed effect?
- We have two or more fixed effects?
- We have too many variables?

In this presentation we present a technique proposed in Carneiro, Guimaraes and Portugal (2010) to estimate a model with 3 high-dimensional fixed effects.
This estimation strategy is discussed in more detail in Guimaraes and Portugal (2010).

The Linear Regression Model

- Consider the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$
- Minimization of the sum of squares (SS) results in a set of equations:

$$\begin{bmatrix} \frac{\partial SS}{\partial \beta_1} = \sum_i x_{1i}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki}) = 0 \\ \frac{\partial SS}{\partial \beta_2} = \sum_i x_{2i}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki}) = 0 \\ \ldots \\ \frac{\partial SS}{\partial \beta_k} = \sum_i x_{ki}(y_i - \beta_1 x_{1i} - \beta_2 x_{2i} - \ldots - \beta_k x_{ki}) = 0 \end{bmatrix}$$

- These equations can easily be solved using

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$$

The Linear Regression Model

- An alternative approach: the partitioned ("cyclic-ascent" or "zigzag") algorithm:
  - 1. Initialize $\beta_1^{(0)}, \beta_2^{(0)}, ..., \beta_k^{(0)}$
  - 2. Solve for $\beta_1^{(1)}$ as the solution to $\frac{\partial SS}{\partial \beta_1} = \sum_i x_{1i}(y_i - \beta_1 x_{1i} - \beta_2^{(0)} x_{2i} - ... - \beta_k^{(0)} x_{ki}) = 0$
  - 2. Solve for $\beta_2^{(1)}$ as the solution to $\frac{\partial SS}{\partial \beta_2} = \sum_i x_{2i}(y_i - \beta_1^{(1)} x_{1i} - \beta_2 x_{2i} - ... - \beta_k^{(0)} x_{ki}) = 0$
  - 3. and so on...
  - 4. Repeat until convergence.

The Linear Regression Model - One Fixed Effect

- Suppose we have a fixed effect: $\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}\alpha + \epsilon$
- where $\mathbf{X}$ is $n \times k$ and $\mathbf{D}$ is a $n \times G_1$ matrix of "dummies" and G_1 is a large number.
- The normal equations are:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{D} \\ \mathbf{D}'\mathbf{X} & \mathbf{D}'\mathbf{D} \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{D}'\mathbf{Y} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X}\beta + \mathbf{X}'\mathbf{D}\alpha = \mathbf{X}'\mathbf{Y} \\ \mathbf{D}'\mathbf{X}\beta + \mathbf{D}'\mathbf{D}\alpha = \mathbf{D}'\mathbf{Y} \end{bmatrix}$$

$$\begin{bmatrix} \beta = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y} - \mathbf{D}\alpha\right) \\ \alpha = \left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'\left(\mathbf{Y} - \mathbf{X}\beta\right) \end{bmatrix}$$

The Linear Regression Model - One Fixed Effect

- This suggests the following "zigzag" estimation procedure:

$$\begin{bmatrix} \beta^{(j+1)} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{Y}-\mathbf{D}\alpha^{(j)}\right) \\ \alpha^{(j)} = \left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'\left(\mathbf{Y}-\mathbf{X}\beta^{(j)}\right) \end{bmatrix}$$

- $\mathbf{D}\alpha$ has dimension $n \times 1$.
- $\left(\mathbf{D}'\mathbf{D}\right)^{-1}\mathbf{D}'$ are group means.
- The "zigzag" approach involves running several regressions with k explanatory variables (1st equation) and repeatedly computing means of residuals (2nd equation).
- The vector $\mathbf{D}\alpha$ contains the estimated fixed effects and if added as a regressor will give the same SS as in a model with the fixed-effects.

Examples

- Estimation of linear regression model with one fixed effect
  EXAMPLE1
- Estimation of linear regression model with one fixed effect (faster approach)
  EXAMPLE2

The Linear Regression Model - Two Fixed Effects

- Suppose we have two fixed effects:

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

- $\mathbf{D}_1$ is $n \times G_1$ and $\mathbf{D}_2$ is $n \times G_2$ and both G_1 and G_2 are large numbers.
- Estimation of this model is complicated. See Abowd, Kramarz and Margolis (Ectrca 1999).
- But a "zigzag" approach is simple to implement:

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

Examples

Estimation of High-Dimensional Models

Paulo Guimarães

Outline

Introduction

LRM - 1FE

LRM - 2FE

LRM - 3FE

Degrees of Freedom

Nonlinear Models

- Estimation of Linear Regression with two fixed effects
  See EXAMPLE3

The Linear Regression Model - Two Fixed Effects

- In practical applications it may make more sense to estimate in steps using the Frisch-Waugh-Lovell theorem.
  - First remove the effects of $\mathbf{D}_1$ and $\mathbf{D}_2$ from $\mathbf{Y}$ and $\mathbf{X}$.
  - Then regress the transformed $\mathbf{Y}$ on the transformed $\mathbf{X}$ to obtain the estimates for β.
  - Then (if needed) recover the estimates of the fixed effects by regressing $\mathbf{u} = \mathbf{Y} - \mathbf{X}\beta$ on $\mathbf{D}_1$ and $\mathbf{D}_2$.
- Regressions on $\mathbf{D}_1$ and $\mathbf{D}_2$ are fast because they only require computation of means.
- We can sweep out one of the fixed effects by demeaning the variables.

Examples

Estimation of High-Dimensional Models

Paulo Guimarães

Outline

Introduction

LRM - 1FE

LRM - 2FE

LRM - 3FE

Degrees of Freedom

Nonlinear Models

- Estimation of Linear Regression with two fixed effects (faster approach)
  See EXAMPLE4

Estimation of the Standard Errors

- The conventional OLS formula

$$V(\widehat{\beta}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

poses 2 problems:

  - How to avoid calculation of $\left(\mathbf{X}'\mathbf{X}\right)^{-1}$
  - How to estimate σ^2 (the problem are the degrees of freedom!)

- The first problem can be solved using

$$V(\widehat{\beta}_j) = \frac{\sigma^2}{Ns_j^2(1-R^2_{j.123\ldots})}$$

- A easier solution is to estimate in two steps. Standard errors (whether or not clustered) of the second equation are correct provided we adjust the degrees of freedom.

Examples

- Estimation of standard errors in the Linear Regression with two fixed effects
  See EXAMPLE5

Stata Commands for LRM with 2 fixed effects

There are 4 user written commands

- `areg` - programmed by Amine Ouazad. Implements the exact least squares solution proposed by Abowd, Creecy and Kramarz (2002). Does not compute standard errors.
- `felsdvreg` - programmed by Thomas Cornelissen. Uses a "memory-saving" approach.
- `gpreg` - programmed by Johannes F. Schmieder. Implements the two-step approach of Guimaraes and Portugal (2009). Some options implemented in Mata. Does not compute clustered standard errors.
- `reg2hdfe` - programmed by Paulo Guimaraes. Implements the two-step approach of Guimaraes and Portugal (2009). Recommended for very large data sets.

reg2hdfe

Command Syntax:

`reg2hdfe` depvar indepvars [if] [in], id1(varname) id2(varname) [options]

where options are:

`fe1`(new varname) `fe2`(new varname)
`cluster`(varname)
`groupid`(new varname)
`outdata`(string)
`maxiter`(integer)
`tolerance`(float)
`indata`(string) To be used after `outdata`(string)
`improve`(string) To be used after `outdata`(string)
`simple check nodots verbose`

Examples

- Estimation of Linear Regression with two fixed effects using reg2hdfe
  See EXAMPLE6

More than two fixed-effects

- Extensions to 3 or more FEs are straightforward
- The normal equations suggest the algorithm

$$\left[\begin{array}{c} \beta = \left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\left(\mathbf{Y}-\mathbf{D}_1\alpha-\mathbf{D}_2\gamma-\mathbf{D}_3\eta\right) \\ \alpha = \left(\mathbf{D}_1'\mathbf{D}_1\right)^{-1}\mathbf{D}_1'\left(\mathbf{Y}-\mathbf{Z}\beta-\mathbf{D}_2\gamma-\mathbf{D}_3\eta\right) \\ \gamma = \left(\mathbf{D}_2'\mathbf{D}_2\right)^{-1}\mathbf{D}_2'\left(\mathbf{Y}-\mathbf{Z}\beta-\mathbf{D}_1\alpha-\mathbf{D}_3\eta\right) \\ \eta = \left(\mathbf{D}_3'\mathbf{D}_3\right)^{-1}\mathbf{D}_3'\left(\mathbf{Y}-\mathbf{Z}\beta-\mathbf{D}_1\alpha-\mathbf{D}_2\gamma\right) \end{array}\right]$$

- Since we can sweep one fixed effect the algorithm should work for 4 FE!
- The only problem is the calculation of degrees of freedom
- In Carneiro et al (2010) we estimate a wage equation with 3 FEs (aprox. 6.4 million workers, 620,000 firms and 115,000 jobs)

Examples

- Estimation of Linear Regression with 3 fixed effects
  See EXAMPLE7

Identification of the fixed effects

- Consider a regression model with N observations and a single fixed effect with G_1 levels (one-way ANOVA model):

$$E(y_{it}) = \mu + \alpha_i$$

- If we replace $E(y_{it})$ by the data cell-means we have a system of G_1 equations on $G_1 + 1$ unknowns
- To solve this model we need to impose one restriction (typically $\mu = 0$ or $\alpha_1 = 0$)
- With this restriction we are able to estimate G_1 coefficients of the model
- This means that SSR has $N - G_1$ degrees of freedom (or $N - k - G_1$ if there are an additional k non-collinear explanatory variables in the model)

Identification of the fixed effects

- Consider now a regression model with two fixed effects with G_1 and G_2 levels respectively,

$$E(y_{it}) = \mu + \alpha_i + \eta_j$$

- Unique combinations of α_i and η_j define a set of equations. We need at least two restrictions to identify the coefficients.
- Consider an example with $G_1 = G_2 = 3$

$$\begin{aligned}
&\mu + \alpha_1 + \eta_1 \\
&\mu + \alpha_1 + \eta_2 \\
&\mu + \alpha_2 + \eta_1 \\
&\mu + \alpha_2 + \eta_3 \\
&\mu + \alpha_3 + \eta_2 \\
&\mu + \alpha_3 + \eta_3
\end{aligned}$$

Identification of the fixed effects

- Impose the restrictions $\mu = 0$ and $\alpha_1 = 0$

$$
\begin{array}{c}
\boxed{\eta_1} \\
\boxed{\eta_2} \\
\alpha_2 + \boxed{\eta_1} \\
\alpha_2 + \eta_3 \\
\alpha_3 + \boxed{\eta_2} \\
\alpha_3 + \eta_3
\end{array}
\rightarrow
\begin{array}{c}
\boxed{\eta_1} \\
\boxed{\eta_2} \\
\boxed{\alpha_2} + \boxed{\eta_1} \\
\boxed{\alpha_2} + \eta_3 \\
\boxed{\alpha_3} + \boxed{\eta_2} \\
\boxed{\alpha_3} + \eta_3
\end{array}
\rightarrow
\begin{array}{c}
\boxed{\eta_1} \\
\boxed{\eta_2} \\
\boxed{\alpha_2} + \boxed{\eta_1} \\
\boxed{\alpha_2} + \boxed{\eta_3} \\
\boxed{\alpha_3} + \boxed{\eta_2} \\
\boxed{\alpha_3} + \boxed{\eta_3}
\end{array}
$$

- The SSR has $N - (G_1 + G_2 - 1) = N - G_1 - G_2 + 1$ degrees of freedom

Identification of the fixed effects

- Impose the restrictions $\mu = 0$ and $\alpha_1 = 0$

$$\begin{array}{ccccc} \mu+\alpha_1+\eta_1 & & \boxed{\eta_1} & & \boxed{\eta_1} \\ \mu+\alpha_1+\eta_2 & & \boxed{\eta_2} & & \boxed{\eta_2} \\ \mu+\alpha_2+\eta_1 & \rightarrow & \alpha_2+\boxed{\eta_1} & \rightarrow & \boxed{\alpha_2}+\boxed{\eta_1} \\ \mu+\alpha_2+\eta_2 & & \alpha_2+\boxed{\eta_2} & & \boxed{\alpha_2}+\boxed{\eta_2} \\ \mu+\alpha_3+\eta_3 & & \alpha_3+\eta_3 & & \alpha_3+\eta_3 \\ \mu+\alpha_3+\eta_3 & & \alpha_3+\eta_3 & & \alpha_3+\eta_3 \end{array}$$

- We would need an additional restriction ($\alpha_3 = 0$ or $\eta_3 = 0$)
- The SSR has $N - (G_1 + G_2 - 2) = N - G_1 - G_2 + 2$ degrees of freedom
- According to Abowd et al 2002 there are now 2 "mobility groups"

Nonlinear models

- This approach can be extended to non-linear models.
- An example with Poisson regression:

$$E(y_i) = \lambda_i = \exp(\mathbf{x}_i'\beta + \alpha_1 d_{1i} + \alpha_2 d_{2i} + \ldots + \alpha_J d_{Ji})$$

- Using the first order conditions:

$$\exp(\alpha_j) = \mathbf{d}_j'\mathbf{y} \times [\mathbf{d}_j' \exp(\mathbf{x}_i'\beta)]^{-1}$$

- Optimization of the maximum-likelihood function requires recursive estimation of a Poisson regression with the **x** variables and an offset containing the estimates α obtained from the expression above.
- The algorithm should work well with models that have globally concave log-likelihood functions

Examples

- A Poisson regression with one fixed effect
  see EXAMPLE8
- A Poisson regression with two fixed effects
  see EXAMPLE9
- A Negative Binomial regression with one fixed effect
  see EXAMPLE10
- A Negative Binomial regression with two fixed effects
  see EXAMPLE11

Final Remarks

- This approach does not require much memory
- It may be extended to many different types of models
- The algorithm is slow but there is room for improvement
- The presentation is based on Guimaraes and Portugal (2010) "A simple feasible alternative procedure to estimate models with high-dimensional fixed-effects"