Estimation of linear panel data models using GMM

Gustavo A. García

ggarci24@eafit.edu.co

Econometría avanzada II

PhD/Maestría en Economía

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

Outline

- Motivation
- General model
- What is GMM?
- GMM and instrumental variables
- System two-stage least squares estimator
- Optimal weight matrix
- Application using R

Lecturas

- Wooldridge, J. (2010). *Econometric Analysis of Cross Section and Panel Data*. 2a edición. MA: MIT Press. Cap 11.1

Motivation

- Use of panel data regression methods has become increasingly popular as the availability of longitudinal data sets has grown
- Panel data contains repeated time-series observations (T) for a large number (N) of cross-sectional units (e.g., individuals, households, or firms)
- An important advantage of using such data is that they allow researchers to control for unobservable heterogeneity, that is, systematic differences across cross-sectional units
- Omitting this unobservable heterogeneity in regressions on aggregated time-series and pure cross-section data, the statistical inferences could be seriously biased

Motivation

- When panel data are available, error-components models can be used to control for these individual differences $\Longrightarrow$ these model assumes that the stochastic error term has two components:
  - a time-invariant individual effect which captures the unobservable individual heterogeneity
  - a usual random noise term
- Some explanatory variables (e.g., years of schooling in the earnings equation) are likely to be correlated with the individual effects (e.g., unobservable talent or IQ). A simple treatment to this problem is the within estimator which is equivalent to least squares after transformation of the data to deviations from means

Motivation

Unfortunately, the within method has two serious defects:

- the within transformation of a model wipes out time invariant regressors as well as the individual effect, so that it is not possible to estimate the effects of time-invariant regressors on the dependent variable
- consistency of the within estimator requires that all the regressors in a given model be strictly exogenous with respect to the random noise (e.g. dynamic models including lagged dependent variables as regressors)

In response to these problems, a number of studies have developed alternative GMM estimation methods. There are several different types of models:

- the linear regression model with strictly or weakly exogenous regressors
- the simultaneous regression model
- the dynamic linear model containing a lagged dependent variable as a regressor

In each case, different assumptions about the exogeneity of the explanatory variables generate different sets of moment conditions that can be used in estimation

General model

We have the following commom model:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + v_{it}; v_{it} = c_i + u_{it} \tag{1}$$

$i = 1, \ldots, N$ represents the cross-sectional unit (individual) and $t = 1, \ldots, T$ indexes time. y_{it} is the dependent variables; $\mathbf{x}_{it} = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{iT})$ is a 1 x K vector of explanatory variables and $\boldsymbol{\beta}$ is a K x 1 vector of unknown parameters

Writing the model (1) for all T time periods as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i \tag{2}$$

donde $\mathbf{v}_i = c_i\mathbf{j}_T + \mathbf{u}_i$
$\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$ is a T x 1 vector
$\mathbf{X}_i = (\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \ldots, \mathbf{x}'_{iT})'$ is a T x K matrix
$\mathbf{v}_i = (v_{i1}, v_{i2}, \ldots, v_{iT})'$ is a T x 1 vector
$\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iT})'$ is a T x 1 vector
$\mathbf{j}_T = (1, 1, \ldots, 1)'$ is the T x 1 vector of ones

General model

There are two well-known special cases of this model: the traditional random effects and fixed effects models. Both of these models assume that the regressors $\mathbf{X}_i$ are strictly exogenous with respect to the random noise u_{it} (i.e., $E(\mathbf{x}'_{it}u_{it}) = \mathbf{0}$ for any t and s)

The random effects model (Balastra and Nerlove, 1966) This model treats the individual effects are as random unobservable which are uncorrelated with all of the regressors

Under this assumption, the parameter vector $\boldsymbol{\beta}$ can be consistently and efficiently estimated by FGLS of the following form:

$$\boldsymbol{\beta}_{RE} = \left(\sum_{i=1}^{N} \mathbf{X}'_i \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{X}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}'_i \widehat{\boldsymbol{\Omega}}^{-1} \mathbf{y}_i\right) \tag{3}$$

where $\widehat{\boldsymbol{\Omega}} = \widehat{\sigma}_u^2 \mathbf{I}_T + \widehat{\sigma}_c^2 \mathbf{j}_T \mathbf{j}'_T$ is the variance matrix estimated of $\mathbf{v}_i$

General model

The fixed effects model

When we treat the c_i as nuisance parameters, the model reduces to the traditional fixed effects model. A simple treatment of the fixed effects model is to remove the effects by the (within) transformation of the model to deviations from individual means:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + u_{it} - \bar{u}_i$$

or

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\boldsymbol{\beta} + \ddot{u}_{it} \tag{4}$$

Writing equation (4) for all time periods as

$$\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\boldsymbol{\beta} + \ddot{\mathbf{u}}_i$$

or

$$\mathbf{Q}_T\mathbf{y}_i = \mathbf{Q}_T\mathbf{X}_i\boldsymbol{\beta} + \mathbf{Q}_T\mathbf{u}_i \tag{5}$$

where $\mathbf{Q}_T = \mathbf{I}_T - \mathbf{j}_t(\mathbf{j}_t'\mathbf{j}_t)^{-1}\mathbf{j}_t'$ is a time-demeaning matrix which is a T x T symmetric, idempotent matrix

General model

Least squares on (5) yields the familiar within estimator:

$$\boldsymbol{\beta}_{FE} = \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i \right)^{-1} \left(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i' \ddot{\mathbf{y}}_i \right) \tag{6}$$

- Although the fixed effects model views the effects c_i as nuisance parameters rather than random variables, the fixed effects treatment (within estimation) is not inconsistent with the random effects assumption

- Mundlak (1978) considers an alternative random effects model in which the effects c_i are allowed to be correlated with all of the regressors $\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$. For this model, Mundlak shows that the within estimator is an efficient GLS estimator

- This finding implies that the core difference between the random and fixed effects models is not whether the effects are literally random or nuisance parameters, but whether the effects are correlated or uncorrelated with the regressors

What is GMM?

- The generalize method of moments (GMM) is a general framework for deriving estimators
- GMM estimators use assumptions about the moments of the random variables to derive an objective function
- The assumed moments of the random variables provide population moment conditions
- The data are used to compute the analogous sample moment conditions
- The parameters estimates are obtained by finding the parameters that make the sample moment conditions as true as possible. This step is implemented by minimizing an objective function
- GMM produces estimators using few assumptions, but is less efficient than, for example, ML that uses the entire distribution while GMM only uses specified moments

What is GMM?

What is generalized about GMM?

- In the method of moments (MM), we have the same number of sample moment conditions as we have parameters

- In the generalized method of moments (GMM), we have more sample moment conditions than we have parameters

What is GMM?

OLS is a MM estimator

We know that OLS estimates the parameters of the conditional expectation of (in cross sectional data) $y_i = \mathbf{x}_i\boldsymbol{\beta} + \epsilon_i$ under the assumption that $E(\epsilon|\mathbf{x}_i) = \mathbf{0}$

Standard probability theory implies that

$$E(\epsilon|\mathbf{x}_i) = \mathbf{0} \implies E(\mathbf{x}_i'\epsilon) = \mathbf{0}$$

So that the population moment conditions for OLS are

$$E[\mathbf{x}_i'(y_i - \mathbf{x}_i\boldsymbol{\beta})] = \mathbf{0}$$

The corresponding sample moment conditions are

$$N^{-1}\sum_{i=1}^{N} \mathbf{x}_i'(y_i - \mathbf{x}_i\widehat{\boldsymbol{\beta}}) = \mathbf{0}$$

Solving for $\widehat{\boldsymbol{\beta}}$ yields

$$\widehat{\boldsymbol{\beta}}_{OLS} = \left(\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{N} \mathbf{x}_i'y_i\right)$$

What is GMM?

OLS is a MM estimator

- The MM only works when the number of moment conditions equals the number of parameters to estimate
- If there are more moment conditions than parameters, the system of equations is algebraically over identified and cannot be solved
- GMM estimators choose the estimates that minimize a quadratic form of the moment conditions
- GMM gets as close to solving the over-identified system as possible
- GMM reduces to MM when the number of parameters equals the number of moment conditions

What is GMM?

Definition of GMM estimator

- Our research question implies q population moment conditions

$$E[\mathbf{m}(\mathbf{w}_i, \boldsymbol{\theta})] = \mathbf{0}$$

$\mathbf{m}$ is q x 1 vector of functions whose expected values are zero in the population
$\mathbf{w}_i$ is the data on person i
$\boldsymbol{\theta}$ is K x 1 vector of parameters, $K \leq q$

- The sample moments that correspond to the population moments are

$$\overline{\mathbf{m}}(\widehat{\boldsymbol{\theta}}) = N^{-1} \sum_{i=1}^{N} \mathbf{m}(\mathbf{w}_i, \widehat{\boldsymbol{\theta}})$$

- When $K < q$, the GMM chooses the parameters that are as close as possible to solving the over-identified system of moment conditions

$$\widehat{\boldsymbol{\theta}}_{\text{GMM}} = \arg\min_{\widehat{\boldsymbol{\theta}}} \overline{\mathbf{m}}(\widehat{\boldsymbol{\theta}})' \mathbf{W} \overline{\mathbf{m}}(\widehat{\boldsymbol{\theta}})$$

GMM and instrumental variables

- We now examine GMM and other related instrumental variables estimators for equation (2)
- Our main focus is a general treatment of given moment conditions, so that we do not make any specific exogeneity assumption regarding the regressors $\mathbf{X}_i$
- We simply begin by assuming that there is a set of T x L instruments $\mathbf{Z}_i$, which satisfies the following orthogonality condition and rank condition

$$E(\mathbf{Z}_i'\mathbf{v}_i) = \mathbf{0} \tag{7}$$

$$\text{rank } E(\mathbf{Z}_i'\mathbf{X}_i) = K \tag{8}$$

Since $E(\mathbf{Z}_i'\mathbf{X}_i)$ is an L x K matrix, rank assumption requires the columns of this matrix to be linearly independent. Necessary for the rank condition is the order condition: $L \geq K$

GMM and instrumental variables

- The orthogonality conditions suggest an estimation strategy
- Under assumptions of orthogonality and rank, $\boldsymbol{\beta}$ is the unique K x 1 vector solving the linear set population moment conditions

$$E[\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})] = \mathbf{0} \tag{9}$$

- Because sample averages are consistent estimators of population moments, the analogy principle applied to condition (9) suggests choosing the estimator $\widehat{\boldsymbol{\beta}}$ to solve

$$N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}}) = \mathbf{0} \tag{10}$$

Equation (10) is a set of L linear equations in the K unknowns in $\widehat{\boldsymbol{\beta}}$

GMM and instrumental variables

Considers the following cases in the solution of the system:

$L = K$: there are exactly enough IVs for the explanatory variables (system exactly identified). Then we can solve for $\widehat{\boldsymbol{\beta}}$ as

$$\widehat{\boldsymbol{\beta}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{X}_i \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{y}_i \right) \tag{11}$$

or

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

$\mathbf{Z}$ is a NT x L matrix; $\mathbf{X}$ is a NT x K matrix; and $\mathbf{Y}$ is a NT x 1 vector

Equation (11) is called the system IV (SIV) estimator

GMM and instrumental variables

$L > K$: more columns in IV matrix $\mathbf{Z}_i$ than we need for identification (system overidentified). Choosing $\widehat{\boldsymbol{\beta}}$ is more complicated and the equation will not have a solution

Instead, we choose $\widehat{\boldsymbol{\beta}}$ to make the vector in equation (10) as "small" as possible in the sample. One possibility is to minimize the squared Euclidean length of the L x 1 vector in equation (10)

This approach suggest choosing $\widehat{\boldsymbol{\beta}}$ to make as small as possible the following expression (dropping N^{-1} in (10)):

$$\left[\sum_{i=1}^{N} \mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})\right]' \left[\sum_{i=1}^{N} \mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\widehat{\boldsymbol{\beta}})\right]$$

While this method produces a consistent estimator under orthogonality and rank assumptions, it rarely produces the best estimator

GMM and instrumental variables

A more general class of estimator is obtained by using a weighting matrix in the quadratic form. Let $\widehat{\mathbf{W}}$ be an L x L symmetric, positive semidefinite matrix, where the "hat" is included to emphasize that $\widehat{\mathbf{W}}$ is a estimator

A generalized method of moments (GMM) estimator of $\boldsymbol{\beta}$ is a vector $\widehat{\boldsymbol{\beta}}$ that solves the problem

$$\min_{\mathbf{b}} \left[\sum_{i=1}^{N} \mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right]' \widehat{\mathbf{W}} \left[\sum_{i=1}^{N} \mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\mathbf{b})\right] \tag{12}$$

The unique solution to the latter expression is

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Z}\widehat{\mathbf{W}}\mathbf{Z}'\mathbf{Y}) \tag{13}$$

System two-stage least squares estimator

A choice of $\widehat{\mathbf{W}}$ that leads to a usual and familiar-looking estimator is

$$\widehat{\mathbf{W}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1} = (\mathbf{Z}'\mathbf{Z}/N)^{-1} \tag{14}$$

When we plug equation (14) into equation (13) and cancel N everywhere, we get

$$\widehat{\boldsymbol{\beta}} = \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X} \right]^{-1} \left[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \right] \tag{15}$$

This is the system 2SLS (S2SLS) estimator

Optimal weight matrix

- Given that a GMM estimator exists for any positive definite weighting matrix, it is important to have a way of choosing among all of the possibilities
- It turns out that there is a choice of $\mathbf{W}$ that produces the GMM estimator with the smallest asymptotic variance, it is achieved when

$$\mathbf{W} = \mathbf{\Lambda}^{-1}$$

$$\mathbf{\Lambda} = \mathrm{Var}(\mathbf{Z}_i'\mathbf{u}_i) = E(\mathbf{Z}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{Z}_i)$$

- Procedure
  1. Let $\check{\boldsymbol{\beta}}$ be an initial consistent estimator of $\boldsymbol{\beta}$. In most cases this is the system 2SLS estimator
  2. Obtain the T x 1 residual vectors $\check{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\check{\boldsymbol{\beta}}$
  3. A generally consistent estimator of $\mathbf{\Lambda}$ is $\widehat{\mathbf{\Lambda}} = N^{-1}\sum_{i=1}^{N} \mathbf{Z}_i'\check{\mathbf{u}}_i\check{\mathbf{u}}_i'\mathbf{Z}_i$
  4. Choose

$$\widehat{\mathbf{W}} = \left(N^{-1}\sum_{i=1}^{N} \mathbf{Z}_i'\check{\mathbf{u}}_i\check{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}$$

     and use this matrix to obtain the asymptotically optimal GMM estimator

- The GMM estimator using the optimal weighting matrix is called the minimum chi-square estimator: $\widehat{\boldsymbol{\beta}}$ is chosen to make the minimum of the objective function (12) that has an asymptotic chi-square distribution

Application using R

We use data from a subsample of the NLSY data on young women aged 14-26 years in 1968. We have a longitudinal dataset of women surveyed during the period 1968-1988 (with gaps). The data is saved in the file nlswork.dta

In the follow links you can found data, data description and code used in R:

- Data
- Data description
- Code in R

Application using R

User Library

```r
library(haven); library(plm); library(tidyverse); library(summarytools); library(AER); library(gmm); library(modelsummary); library(gt); library(DT)
```

Reading data and processing information. We have 4711 women employed, not enrrolled in school and who have completed their education

```r
nlswork <- read_dta("http://www.stata-press.com/data/r10/nlswork.dta") %>% # leemos la base de datos
  mutate(age2 = age*age) %>%      # construimos la variables age2
  select(idcode, year, ln_wage, age, age2, tenure, union, not_smsa, south) |> # seleccionando var
  filter(union !='NA', south !='NA')
```

Set panel structure

```r
nlswork <- pdata.frame(nlswork, c("idcode","year"))
datatable(head(nlswork, 5))
```

Show 10 entries Search:

	idcode	year	ln_wage	age	age2	tenure	union	not_smsa	south
1-72	1	72	1.58997738361359	20	400	0.916666686534882	1	0	0
1-77	1	77	1.77868056297302	25	625	1.5	0	0	0
1-80	1	80	2.55171537399292	28	784	1.83333337306976	1	0	0
1-83	1	83	2.42026138305664	31	961	0.666666686534882	1	0	0
1-85	1	85	2.61417245864868	33	1089	1.91666662693024	1	0	0

Application using R

Regressions IV in datapanel

```
# OLS
ols_pool <- lm(ln_wage ~ tenure+age+I(age2)+not_smsa, data=nlswork)

# IV1
iv1_pool <- ivreg(ln_wage ~ tenure+age+I(age2)+not_smsa |
                      union+south+age+I(age2)+not_smsa, data=nlswork)

# IV2
iv2_pool <- tsls(ln_wage ~ tenure+age+I(age2)+not_smsa,
                         ~  union+south+age+I(age2)+not_smsa, data=nlswork)

# IV3
iv3_pool <- plm(ln_wage ~ tenure + age + I(age2) + not_smsa |
              union + south + age + I(age2) + not_smsa, data = nlswork, model = "pooling")

# GMM with vcov="iid"
gmm_iid <- gmm(ln_wage ~ tenure+age+I(age2)+not_smsa,
           ~  union+south+age+I(age2)+not_smsa, data=nlswork, vcov="iid")

# GMM with vcov="HAC"
gmm_hac <- gmm(ln_wage ~ tenure+age+I(age2)+not_smsa,
               ~  union+south+age+I(age2)+not_smsa, data=nlswork, vcov="HAC")

models <- list("OLS" = lm(ln_wage ~ tenure+age+I(age2)+not_smsa, data=nlswork),
               "IV1" = ivreg(ln_wage ~ tenure+age+I(age2)+not_smsa |
                               union+south+age+I(age2)+not_smsa, data=nlswork),
               "IV2" = tsls(ln_wage ~ tenure+age+I(age2)+not_smsa,
                            ~  union+south+age+I(age2)+not_smsa, data=nlswork),
               "IV3" = plm(ln_wage ~ tenure + age + I(age2) + not_smsa |
                             union + south + age + I(age2) + not_smsa, data = nlswork, model = "pooling"),
               "GMM iid" = gmm(ln_wage ~ tenure+age+I(age2)+not_smsa,
                           ~  union+south+age+I(age2)+not_smsa, data=nlswork, vcov="iid"),
               "GMM hac" = gmm(ln_wage ~ tenure+age+I(age2)+not_smsa,
                           ~  union+south+age+I(age2)+not_smsa, data=nlswork, vcov="HAC"))
```

Application using R

```
modelsummary(models, output="gt", coef_map = c('tenure' = 'Tenure', 'age' = 'Age', 'I(age2)' = 'Age2', 'not_smsa' = 'Not SMSA (=1)', '(Intercept)' = '
            estimate = "{estimate}{stars}", title = "Table 1. Effects of tenure on wages", gof_omit='IC|Log|F|R2', stars = c('*'=.1,'**'=.05,'***'=.(
  tab_style(style = cell_text(size = 'small'), locations = cells_body()) |>
  tab_style(style = cell_text(color='blue'), location = cells_body(row=1:2)) |>
  tab_style(style = cell_text(color='black', size='x-small'), locations = cells_source_notes()) |>
  tab_source_note(source_note = "Note: standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01")
```

Table 1. Effects of tenure on wages

	OLS	IV1	IV2	IV3	GMM iid	GMM hac
Tenure	0.037***	0.149***	0.149***	0.149***	0.149***	0.136***
	(0.001)	(0.007)	(0.007)	(0.007)	(0.007)	(0.011)
Age	0.047***	0.025***	0.025***	0.025***	0.025***	0.026***
	(0.005)	(0.007)	(0.007)	(0.007)	(0.007)	(0.010)
Age2	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Not SMSA (=1)	-0.251***	-0.249***	-0.249***	-0.249***	-0.249***	-0.243***
	(0.007)	(0.010)	(0.010)	(0.010)	(0.010)	(0.017)
Constant	0.901***	1.209***	1.209***	1.209***	1.209***	1.189***
	(0.070)	(0.102)	(0.102)	(0.102)	(0.102)	(0.136)
Num.Obs.	19007	19007	19007	19007	19007	19007

Note: standard errors in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01