TA Session 1

Microeconometrics with Hanna Wang
IDEA, Fall 2022

TA: Conghan Zheng

TA sessions

- 10 sessions, 1 hour each (Fridays 3pm-4pm).
- A week before the next PS is due, we go over exercises that prepare you for the problem sets.
- A week after the last PS is due, we also talk about common mistakes in your solutions, if any.
- Office hours: Wednesday afternoons 3pm-6pm, better send me a email before to confirm.

Problem sets

- 6 problem sets, one for each chapter.
- The deadlines are Fridays the minute my session starts, late submission will not be accepted, especially after I posted the solutions.
- You can work in groups, and everyone is expected to submit their own solutions.
- Similar sentences in the interpretations of results will cause you a huge grade loss. Creative and insightful interpretations and discussions will earn you a bonus in the grade of that solution.
- You may get answers from past cohorts, but I suggest you don't look at the answers when you are solving the PS to avoid any anchoring effects. I'm the one who wrote the answers, I will recognize similar interpretations.
- This course recommends you use Stata. For PS solutions I also accept Matlab, R, Python, and Julia (these are all the languages I can give you feedback on).
- When you get stuck on anything in the PS, contact me via email: conghan.zheng@uab.cat

Overview

1. GMM
2. MLE
3. GMM and MLE

GMM

Moments

	SAMPLE	POPULATION
	$\{y_i\}_{i=1}^N$	$f(y, \beta)$
Moments :	$\bar{y}$	$\mathbb{E}(y_i) = \mu$
	$\sum_{i=1}^N y_i^2$	$\mathbb{E}(y_i^2) = \sigma^2 + \mu^2$
	$\vdots$	$\vdots$
Regression functions:	$y = x\hat{\beta} + e$	$y = x\beta + \varepsilon$
	$\vdots$	$\vdots$
		(unobservable)

- The term *regression* is used to signify a predictive relationship between y and x.
- Regression function is just one particular feature of the distribution $f(y, \beta)$. The regression function of y on x: $E(y|x)$ is the first conditional moment.

Moment Conditions

- Matching sample moments with population moments gives us the Moment Conditions.

$$\left.\begin{array}{c} \bar{y} = \mu(\hat{\beta}) \\ \sum_{i=1}^{N} y_i^2 = \sigma^2(\hat{\beta}) + \mu^2(\hat{\beta}) \\ \vdots \end{array}\right\} \Rightarrow \hat{\beta}_{MME} \text{ for true } \beta$$

- Method of moments estimator (MME) solves the moment conditions. For example, the sample mean is the MME of the population mean.

Moment Conditions

- Population moment conditions:

$$\mathbb{E}\left[m(y_i, x_i, z_i, \beta)\right] = 0$$

  - all observables of the i-th individual: x_i ($K \times 1$ vector), y_i (scalar), and instrument z_i ($1 \times L$ vector);
  - unknown parameters: β ($K \times 1$ vector)
  - moment function: m ($L \times 1$ vector)
- Sample moment conditions:

$$\bar{m}(y_i, x_i, z_i, \hat{\beta}) = \frac{1}{N} \sum_{i=1}^{N} m(y_i, x_i, z_i, \hat{\beta}) = 0$$

- (Necessary) Order condition for identification:

$$L \geq K$$

(no. of orthogonality conditions $\geq$ no. of parameters)

GMM

- When $L > K$, there are more moment conditions (hence equations to solve) than there are parameters, MME is infeasible, and thus the GMM is created.
- For example, when you have more instruments than what you are going to instrument for, there is no analytical solution to the system of moment conditions.
- When there is not solution to set $\bar{m}(\hat{\beta})$ exactly equal to zero, we can consider choosing some $\hat{\beta}$ such that $\bar{m}(\hat{\beta})$ is as ***close*** to zero as possible (this is where the *G* in the name *GMM* comes from: generalized).

GMM

- To make precise what we mean by ***close***, we define the ***distance*** between two vectors: $m(\hat{\beta})$ and the zero vector by the quadratic form

$$\left[m(\hat{\beta}) - 0\right]' W^{-1} \left[m(\hat{\beta}) - 0\right],$$

where W is a symmetric and p.s.d. matrix defining the distance.

- By minimizing the quadratic distance from $m(\hat{\beta})$ to zero, we obtain the **generalized method of moments (GMM) estimator** of β^0:

$$\hat{\beta}_{GMM} \equiv \arg\min_{\beta} \; m(\hat{\beta})' W^{-1} m(\hat{\beta})$$

  - When $W = I$, an identity matrix, each of the L sample moment components is weighted equally.
  - When the weight matrix $W \neq I$, the sample moment components are weighted differently. A suitable choice of weighting matrix can improve the efficiency of the resulting estimator.

GMM in practice

- **Question**: How to specify the optimal weighting function?
  - Intuitively, the sample moment components which have large sampling variations should be discounted, which is an idea similar to GLS. An optimal choice of W produces the GMM estimator with the smallest asymptotic variance (highest efficiency).
- **Question**: Does the GMM estimator have a closed form expression?
  - In general, when the moment function $m(w_i, \beta)$ is nonlinear in β, there is no closed form solution for $\hat{\beta}$.
  - Special case: linear IV, an important special case where the GMM estimator has a closed form.

GMM in practice

Estimator	IV	2SLS	Optimal (Efficient) GMM
Rank condition	$L = K$	$L \geq K$	$L \geq K$
Case	independent and homoskedastic errors	independent and homoskedastic errors	independence or homoskedasticity of errors is violated
Weighting matrix	$\hat{W} = z'z$		$\hat{W} = \widehat{Var(\frac{1}{\sqrt{N}} z'\varepsilon)}^{-1} = \left(\frac{1}{N}\sum_{i=1}^{N} e'ez_i z_i'\right)^{-1}$

Expressions of estimators:

$$\beta_{IV} = (z'x)^{-1}z'y,$$

$$\beta_{2SLS} = (x'z(z'z)^{-1}z'x)^{-1}x'z(z'z)^{-1}z'y$$

$$\beta_{OGMM} = (x'z\hat{W}z'x)^{-1}x'z\hat{W}z'y$$

GMM in practice
Optimal GMM

- Steps of 2-stage optimal GMM[1]:
  - Step 1- find a initial consistent estimator $\tilde{\beta}$ for β for some prespecified W which converges in probability to some finite and p.d. matrix. For convenience we set $W = I$, and get a consistent but inefficient estimator:

$$\tilde{\beta} \equiv \arg\min_{\beta} \quad m(\beta)' m(\beta)$$

  - Step 2- use the residual $e = y - x\tilde{\beta}$ form the 1st step obtain a consistent estimator for W: $\hat{W} = \left(\frac{1}{N}\sum_{i=1}^{N} e_i^2 z_i z_i'\right)^2$, and use $\hat{W}$ for the second stage minimization:

$$\hat{\beta}_{GMM} \equiv \arg\min_{\beta} \quad m(\beta)' \hat{W} m(\beta)$$

- In practice, sometimes we need to iterate until the estimate from step 2 converges (iterated GMM).

[1] Linrearity, ergodic stationarity, and finite fourth moments are assumed (knowledge of which is not important to this course).

GMM in practice

`TA1.dta`: an extract from the Medical Expenditure Panel Survey of individuals over the age of 65 years.

- Benchmark model:

$$y = \alpha d + \beta X + \varepsilon$$

- y: variable `ldrugexp` in data, log of total out-of-pocket expenditure
- d: variable `hi_empunion` in data, whether the individual holds either employer- or union-sponsered health insurance, we treat it as endogenous in this example
- X: other controls
- z: available instruments for d
  - `ssiration`: the ratio of an individual's social security income to total income
  - `multlc`: whether the firm is a large operator with multiple locations

GMM in practice

- Just-identified model:
  - use one instrument to instrument for d (`hi_empunion`)

	(1) OLS	(2) IV	(3) GMM_1S	(4) GMM_Optimal	(5) GMM_iterated
hi_empunion	0.0739 (0.0261)	-0.898 (0.208)	-0.898 (0.208)	-0.898 (0.221)	-0.898 (0.221)
totchr	0.440 (0.00957)	0.450 (0.0104)	0.450 (0.0104)	0.450 (0.0102)	0.450 (0.0102)
age	-0.00353 (0.00189)	-0.0132 (0.00287)	-0.0132 (0.00287)	-0.0132 (0.00300)	-0.0132 (0.00300)
female	0.0578 (0.0252)	-0.0204 (0.0315)	-0.0204 (0.0315)	-0.0204 (0.0326)	-0.0204 (0.0326)
blhisp	-0.151 (0.0338)	-0.217 (0.0387)	-0.217 (0.0387)	-0.217 (0.0395)	-0.217 (0.0395)
linc	0.0105 (0.0139)	0.0870 (0.0220)	0.0870 (0.0220)	0.0870 (0.0226)	0.0870 (0.0226)
_cons	5.861 (0.153)	6.787 (0.255)	6.787 (0.255)	6.787 (0.269)	6.787 (0.269)
N	10089	10089	10089	10089	10089

*Standard errors in parentheses

GMM in practice

- Over-identified model:
  - use two instruments to instrument for d (`hi_empunion`)

	(1) IV	(2) GMM_1S	(3) GMM_Optimal	(4) GMM_iterated
hi_empunion	-0.990	-0.990	-0.993	-0.993
	(0.192)	(0.192)	(0.205)	(0.205)
totchr	0.451	0.451	0.451	0.451
	(0.0105)	(0.0105)	(0.0103)	(0.0103)
age	-0.0141	-0.0141	-0.0142	-0.0142
	(0.00278)	(0.00278)	(0.00290)	(0.00290)
female	-0.0278	-0.0278	-0.0282	-0.0282
	(0.0312)	(0.0312)	(0.0322)	(0.0322)
blhisp	-0.224	-0.224	-0.223	-0.223
	(0.0387)	(0.0387)	(0.0396)	(0.0396)
linc	0.0943	0.0943	0.0945	0.0945
	(0.0212)	(0.0212)	(0.0219)	(0.0219)
_cons	6.875	6.875	6.878	6.878
	(0.245)	(0.245)	(0.258)	(0.258)
N	10089	10089	10089	10089

*Standard errors in parentheses

Multiple Equation GMM

Example

- Consider this simultanuous-equation model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_1 \tag{1}$$

$$x_1 = \beta_3 + \beta_4 y + \beta_5 x_3 + \varepsilon_2 \tag{2}$$

- Variables y and x_1 are endogenous.
- Suppose we have available instrument vector z_1 for x_1 in equation 1, vector z_2 to instrument for y in (2)[2].

[2] Refer to Wooldridge(2010) Chapter 9.6 for a discussion of instances where it is necessary to use different instruments in different equations.

Multiple Equation GMM

Example

Separated estimation:

- Stage 1: estimate (1) using GMM.

$$\text{moment condition:} \quad \mathbb{E}[z_1(y - \beta_0 - \beta_1 x_1 - \beta_2 x_2)] = 0$$

$$\text{optimal weight matrix :} \quad \hat{W}_1 = \left(\frac{1}{N} \sum_{i=1}^{N} e_1' e_1 z_1 z_1' \right)^{-1}$$

Notice that e_1 is the residual from the first stage of the optimal GMM.

- Stage 2: estimate (2) using GMM, use the predicted $\hat{y}$ from last step.

$$\text{moment condition:} \quad \mathbb{E}[z_2(x_1 - \beta_3 - \beta_4 \hat{y} - \beta_5 x_3)] = 0$$

$$\text{optimal weight matrix :} \quad \hat{W}_2 = \left(\frac{1}{N} \sum_{i=1}^{N} e_2' e_2 z_2 z_2' \right)^{-1}$$

Notice that e_2 is the residual from the first stage of the optimal GMM.

Multiple Equation GMM

Example

Joint estimation:

- Estimate the system jointly.

$$\text{moment conditions:} \quad \mathbb{E}\begin{Bmatrix} z_1(y-\beta_0-\beta_1x_1-\beta_2x_2) \\ z_2(x_1-\beta_3-\beta_4y-\beta_5x_3) \end{Bmatrix} = 0$$

$$\text{weighting matrix:} \quad \hat{W} = \frac{1}{2N}\sum_{i=1}^{N} e'e \cdot zz'$$

$$\text{where} \quad e = \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} = \begin{pmatrix} y-\hat{\beta}_0-\hat{\beta}_1x_1-\hat{\beta}_2x_2 \\ \\ x_1-\hat{\beta}_3-\hat{\beta}_4y-\hat{\beta}_5x_3 \end{pmatrix}$$

$$z = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

where e is from the first stage of optimal GMM.

Multiple Equation GMM
Example

- See `TA1.do` for the following example. From the following table, joint estimation has lower equation 1 estimates and higher equation 2 estimates.

	(Eq1)	(Eq2)	(Joint)
wagepriv	0.801 (0.0848)		0.778 (0.0661)
wagegovt	1.030 (0.237)		0.975 (0.238)
consump/consump_hat		0.362 (0.239)	0.428 (0.198)
govt		1.144 (0.489)	1.114 (0.388)
capital1		0.0300 (0.0751)	-0.0256 (0.0547)
_cons(Eq1)	19.36 (2.752)		20.50 (2.056)
_cons(Eq2)		5.373 (15.50)	12.84 (11.68)
N	22	22	22

*Standard errors in parentheses

Multiple Equation GMM

Example

- Joint estimation has lower equation 1 estimates and higher equation 2 estimates.
- What causes the difference? Consider the estimates for x_1 (`wagepriv`) and y (`consump`) for example.
  - Estimates of β_1 and β_4 are both positive, x_1 and y reinforce each other through causality (equation (1)) and reverse causality (equation (2)).
  - In separeted estimations, when we estimate equation (1), it's equivalent to that we only take the reduced-form one-way relationship between x_1 and y, the positive feedback effect of y on x_1 (defined in equation (2)) is taken as part of the effects of x_1 on y (equation (1)). More variation is used in the first stage, hence $\hat{\beta}_1$ will be higher than the true value. Less variation is left for the second stage, which leads us to lower equation (2) estimates.
  - Therefore we can say, joint estimates are closer to the true value.

MLE

MLE

- Instead of modelling conditional mean or conditional expectation, is there any case which calls for modelling the conditional probability distribution of y given x?
- Topics in Chapter 3 to 5 of this course.

MLE

- A random sample of size N is a collection of random vectors $Z^N = (Z_1', \dots, Z_N')$, where $Z_i = (Y_i, X_i')', i \in \{1, \dots, N\}$.
- A realization of Z^N is a data set, denoted as $z^N = (z_1', \dots, z_N')$. A random sample can generate many realizations (i.e., datasets).
- **Likelihood function**:

$$L_N(\theta; z^N) = f(z^N, \theta)$$

  where $f(\cdot)$ is the joint probability density/mass function of the random sample.
- The likelihood function is algebraically identical to the joint probability density function of the realized random sample.

FIML and LIML

For a system of simultanuous equations,

- **Full infomation maximum likelihood (FIML)** estimates the complete system jointly. With normal errors, FIML is efficient among all estimators, but it's not commonly used in applied econometrics.
- **Limited information maximum likelihood (LIML)** is an estimator based on sequential estimations of the system (which is less efficient than FIML in cases where they are comparable, for example, nested logit).
- LIML is the ML counterpart of 2SLS (they share the same asymptotic distribution) and is efficient among single equation estimators. FIML is the ML counterpart of 3SLS[3].
- The FIML estimator is not consistent if any part of the system is misspecified, and it always requires more computing power. If you wishes to estimate only one structural equation of interest, LIML is prefered.

[3] 3SLS: GMM estimation of a multiple-equation system, first perform equation-by-equation 2SLS, then compute a weighting matrix on joint moment conditions.

GMM and MLE

GMM and MLE

- GMM estimator can be more robust than MLE
  - MLE requires a fully specified model. GMM requires less assumptions, only limited moment conditions need to be correctly specified, so it's robust to distributional misspecification. The price for robustness is usually a loss of efficiency.
- MLE can be more efficient than GMM estimator
  - MLE would be asymptotically more efficient than the best GMM estimator, but failure of normality generally results in inconsistent estimators of all parameters.
  - MLE uses all the information of the entire distribution and therefore is not always feasible. GMM uses only the specified moments and therefore requires less computational power.
- ML is a special case of GMM
  - The ML score equations can be viewed as moment conditions. GMM estimator with optimal orthogonality conditions is asymptotically equivalent to MLE.

References

- Hayashi, F. (2000). Econometrics. Princeton University Press. Chapter 3, 7, 8.
- Wooldridge, J. M. (2010). Econometric analysis of cross section and panel data. MIT press. Chapter 8.
- Cameron, A. C., & Trivedi, P. K. (2022). Microeconometrics using stata (Second Edition). Stata press. Chapter 7.