607
GMM in relevant sample sizes

Jon Faust

http://e105.org/e607

October 19, 2015

- **Readings**
  - See the GMM section of the course website and links throughout.
- **GMM for economists is like magic.**
- **Magic**
  - Hansen-Singleton, 1982, p.1285

    The advantagesa of these procedures are that they circumvent the need for explicitly deriving decision rules, and they do not require the specification o fthe joint distribution function of the observable variables. The techniques are appropriate for any dynamic model whose econometric implications can be cast in terms of a set of orthogonality conditions.
- **cite**
  - Hansen-Singleton, 1982, Econometrica

    go

    www.jstor.org/stable/1911873
- **Why GMM is wonderful**
  - Our models are inevitably nonlinear
  - GMM can give estimates of general nonlinear models using one simple recipe.

    Grab some moments and minimize a quadratic form
  - ML in contrast would have to be adapted to each likelihood
  - Our models naturally produce interpretable moment conditions

Often in the form of Euler equation errors being conditionally mean zero

- Or (often equivatently) expectations errors for the agents are zero
- Imprecise models. We have models in which we may believe certain key relations, and would like to either leave other aspects (such as details of dynamics) unstated or we state them, but we really don't want to 'take them seriously'
- GMM allows us to focus the estimation on relations we'd like to focus on.
- And the estimates will even be efficient among the class of estimators exploiting those moments.

▶ **Key choices**

- Which moments? And (when relevant) which instruments?
- Which weight matrix?

  Use efficient GMM matrix or some constrained form

- How to estimate weight matrix.

  Two-step or continuous updating; practical choices such as truncation lag.

▶ **These choices matter**

- A large body of work grew up quickly with the spread of GMM showing that small sample properties are often highly sensitive to all these choices

  Despite the fact that these don't matter in large samples.

▶ **Example 1:**

- The first and classic use of GMM is to test the implications of the consumption Euler equation for asset prices.

  Hansen-Singleton article cited above

- Basically, optimization in simple cases implies that for every asset held by the agent,

$$1 = E_t m_{t,t+k} R_{t,t+k}$$

  where R_t is the gross real return on the asset between t and $t+k$ and $m_{t,t+k}$ is known as the pricing kernel and for simple cases is given by

$$m_{t,t+k} = \beta^k u'(c_{t+k})/u'(c_t)$$

▶ **Thus,**

- We form a moment condition

$$0 = E[(m_{t,t+k}R_{t,t+k} - 1) \times z_t]$$

- This relation should hold for the return of every asset held by the agent and for every instrument z_t that is in the agent's info. set at t.
- That is, we can form lots of moment conditions and have zillions of instruments.

▶ **Magic?**

- Kocherlakota runs a Monte Carlo on Hansen-Singleton's method

  JME, 1990. go

  http://www.sciencedirect.com/science/article/pii/030439329090024X

- Finds that the method can be really ill-behaved in relevant cases

▶ **Kocherlakota, teaser result**

- p.298

  All in all, the estimators with multiple instruments perform poorly in small samples of data from the model economy. They underestimate the preference parameters and the standard errors. Morover, they reject the model too often at the 5 percent level of significance.

- Read the paper for some speculation as to why.

▶ **Example 2: Stochastic vol. model**

- Cite: Anderson Sorensen, 1996

  JBES, 1996 go

  http://www.jstor.org/stable/1392446

▶ **DGP**

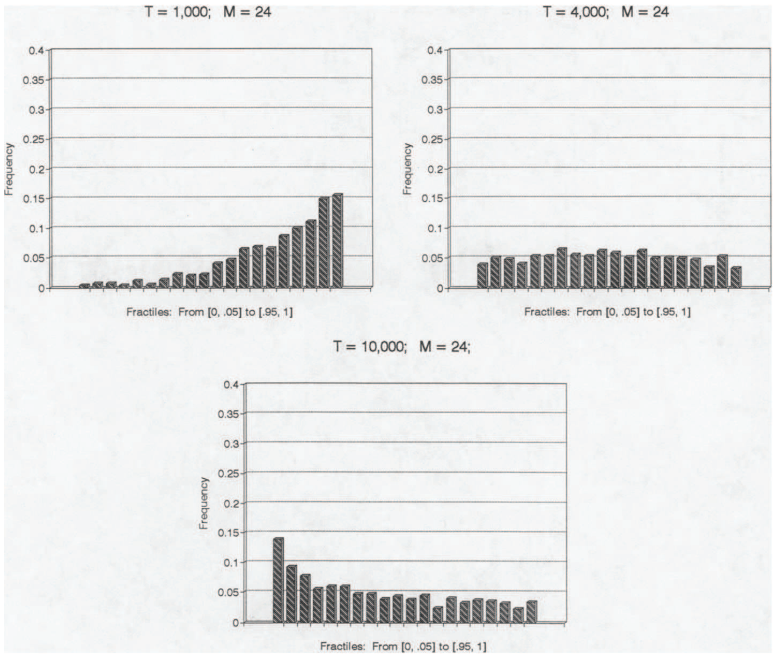
- log-normal SV (stochastic volatility) model

$$\begin{aligned} y_t &= \sigma_t Z_t \\ \log(\sigma_t^2) &= \omega + \beta \ln(\sigma_{t-1}^2) + \sigma_u u_t \end{aligned}$$

  and $(Z_t, u_t) \sim iidN(0, I)$.

- Paramerter: $\theta = (\omega, \beta, \sigma_u)$.
- y_t strictly stationary for $0 < \beta < 1$ and $\sigma_u > 1$.

- **The estimators**
  - ML might be a pain: kind of a pain to work out the implied likelihood from this DGP and to figure out how to maximize it efficiently.
  - Alternative: Pick θ to make the first Q sample moments as close as possible to the implied population moment implied by θ.
    - j^{th} pop. moment: $E_\theta y_t^j$
    - j^{th} sample momente: $T^{-1} \sum y_t^k$
  - Of course, we have various choices for the weight matrix, I, efficient GMM, etc.
- **Monte Carlo**
  - Run a big Monte Carlo with the DGP calibrated to look like real-world data.
- **Long-story short**
  - Some choices for the number of moments and weight matrix perform very badly
  - There are important tradeoffs between moment choice and weight matrx choice
- **One indicative result: background**
  - Remember that the GMM criterion function itself has the form of a standard Wald test
  - The value of the criterion at the max. is known in this context as the J test of the overidentifying restrictions implied by the model.
- **One indicative result**
  - Let's look at the true size of a nominal 5 percent J test based on a particular choice of the weight matrix
  - We'll use the first 24 moments, various sample sizes
- **Anderson Sorensen, Fig. 6**



- **Aside:: Distribution of p-values under the null**
  - The figure summarizes the distribution of p-values under the asymptotic $\chi_{(2)}$ distribution.
  - You should be able to show that under the null hypothesis, p-values in epeated sampling should be uniformly distributed.

- **Size distortion**
  - For $T = 1000$, the we have far too few values in the 0 to 0.05 range.
  - For $T = 10,000$, we have far too many in this range.
  - In short, the true size is not close to the nominal size of the test and the nature of the distortion changes with sample size.

- **Subsidiary lesson**
  - Sometimes a simple choice of W, say $W = I$ dominates the (asymptotically) efficient choice as far as precision of the point estimates is concerned.

- **Example 3: Faust-Wright**
  - Faust and Wright: Efficient prediction of excess returns.

    RESTat, 2011 go

    http://www.mitpressjournals.org/doi/pdf/10.1162/REST%5Fa%5F00092

- Paper is not very significant a contribution, but has a lot of pedagogic value.

  relations among OLS, SUR, GMM, EGMM. Relative efficiency of these. Relevant sample size applicability of the large sample results.

▶ **Goal in the paper**

- Show how to improve estimates of forecastability of excess returns in cases where
  - You think the return is mainly unforecastable
  - But there may be a small forecastable component

▶ **Simple version**

- Simple predictive regression

$$y_t = \beta' x_{t-h} + \varepsilon_t$$

  y_t is return from $t-h$ to t, x_{t-h} is in the agent's info. set at $t-h$.

- Run the regression, see if the βs are all zero.

  If we reject, we have found predictability.

▶ **Motivation**

- Everyone knows that excess returns are highly variable
- And nobody thinks that they are very predictable
- Thus, we are trying to estimate β in a very noisy environment

  variance of ε large relative to that of $\beta' x_t$

- Suppose

$$\varepsilon_t = \phi' w_t + \xi_t$$

- Where ξ_t is purely unpredictable and unobservable
- w_t is unpredictable but measureable ex post

  Thus, you can't use w as a predictor in real time, but w might help you form better estimates of β.

▶ **Huh? What is w_t?**

- We think that financial market returns in response to the 9-11 terrorist attack or or in response to hurrican Katrina were not forecastable from, say, a month in advance.
- That is, some big surprising events drive returns and we think those movements are not predictable in advance.

▶ **Simple case**

- Suppose ϕ is known.
- Just subract $\phi' w_t$ from y_t before running the predictive regression:

$$\tilde{y}_t = x_t' \beta + \xi_t$$

  where $\tilde{y}_t = y_t - \phi' w_t$ is the return purged of the effect of w.
- You should be able to show that the relative efficiency of the original regression and this one is $1 - \lambda$ where λ is the population R^2 from running ε_t on w_t.

  In the scalar w case, the squared correlation of ε_t and w_t.

▶ **ϕ unknown.**

- With ϕ unknown we have a more subtle case
- Faust-Wright show how to cast this as a case of SUR
- And under strong assumptions, the SUR estimator is efficient
- And the SUR estimator happens to correspond to just augmenting the original equation with w_t.

▶ **The augmented regression**

- The augmented regression:

$$y_t = x_{t-h}' \beta + \phi' w_t + \xi_t$$

▶ **But**

- The simplest SUR estimator does not take account of possible serial correlation in the implied moment conditions.
- That, is, SUR can be cast as GMM, but is not efficient GMM under general assumptions.

▶ **Thus, 3 estimators**

- Simple OLS on the original equation

  This is GMM with weight matrix I using only ε as an instrument

- OLS on the augmented equation, which is GMM with an inefficient choice of weight matrix

  This is GMM with instruments including ε and w, but not the efficient weight matrix.

- Efficient GMM

  Fully efficient GMM with instruments ε and w.

▶ **Result**

- A Monte Carlo shows that OLS on the augmented regression seems to perform best in relevant cases
- Efficient GMM can give very erratic results.

▶ **Aside:: In the sample empirical exercise**

- For predicting excess bond returns, the proposed method can give substantial gains relative to simple OLS on the original equation.
- For stock returns (where the predictable component is though to be much smaller than that in bonds) the method doesn't change things much relative to simple OLS.

▶ **Key intuition**

- Define ℓ_t to be the instruments that the x are orthogonal to. This includes *eps* and w.
- The weight matrix for efficient GMM on the augmented regression is a consistent estimate of:

$$\lim ET^{-1} \sum_s \sum_t (x_t x_s' \otimes \ell_t \ell_s')$$

  This is our usual double sum for general cases of variance-covariance matrices

- OLS on the agumented regression uses a consistent estimate of only the $s = t$ terms in the above expression
- The result that the simpler weight matrix works better essentially suggests that estimating all those terms to capture serial correlation ($t \neq s$ terms) add sufficient sloppiness as to make the estimates unreliable in relevant sample sizes.
- Of course, in sufficiently large samples, this cannot be an issue: the consistent estimator used in EGMM will be as precise as you like as T gets large.

▶ **Bottom line: H^3.**

- All three examples show ways in which doing more complicated things (adding moment conditions, using the EGMM weight matrix, etc.) that make your results unambiguously better in large enough sample sizes often make your results erratic and unreliable in smaller samples.
- I sometimes call this H^3: heavy-handedness helps.

  Specifically, using the estimate '0' for lots of stuff (or shrinking estimates strongly toward zero) increases bias, but does wonders for variance. In weakly informative samples, reducing variance may be more important than modest asymptotic efficiency gains.

- A great deal of basic econometric theory is directed toward ways to relax assumptions and, using more sophisticated techniques, get answers that remain efficient.

- This is very important work.
- Much less effort tends to be directed toward helping us wisely decide whether the advanced techniques make us better off or worse off in relevant sample sizes.
- Regularly when we check we find that that elaborate methods decrease reliability of estimates.
- Thus, the KISS and H^3 principles

  Keep it simple, stupid. Heavy-handedness helps.

- The profession is slowly providing a sounder theoretical basis for KISS and H^3 that can guide the applicaiton of these principles
- There is much to do both on the theory side
- And until we have a strong theory basis, there is much to do refining 'good practice'