14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 8, 2007
corrected September 2012

Lecture 6

GMM

This lecture extensively uses lectures given by Jim Stock as a part of mini-course at NBER Summer Institute.

Intro to GMM

We have data z_t, parameter θ, and moment condition $Eg(z_t, \theta_0) = 0$. This moment condition identifies θ. Many problems can be formulated in this way.

Examples

- **OLS**: we have assumption that $y_t = x_t'\beta + e_t$ and $Ee_t x_t = 0$. We can write this as a moment condition, $E[x_t(y_t - x_t'\beta)] = 0$, that is, $g(x_t, y_t, \beta) = x_t(y_t - x_t'\beta)$.
- **IV**: Consider a case when $y_t = x_t'\beta + e_t$, but the error term may be correlated with the regressor. However, we have so called instruments z_t correlated with x_t but not correlated with the error term: $Ee_t z_t = 0$ This gives the moment condition $E[z_t(y_t - x_t'\beta)] = 0$, and $g(x_t, y_t, z_t, \beta) = z_t(y_t - x_t'\beta)$.
- **Euler Equation**: This was the application for which Hansen developed GMM. Suppose we have CRRA utility, $u(c) = \frac{c^{1-\gamma}-1}{1-\gamma}$. An agent solves the problem of smoothing consumption over time (inter-temporal optimization). The first order condition from utility maximization gives us the Euler equation,

  $$E\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} R_{t+1} - 1 | I_t\right] = 0,$$

  where I_t is information available at time t (sigma-algebra of all variables observed before and at t). For any variable z_t measurable with respect to I_t, we have

  $$E\left[\left(\beta\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} R_{t+1} - 1\right) z_t\right] = 0,$$

  This moment condition can be used to estimate γ and β.

Estimation

Take the moment condition and replace it with its sample analog:

$$Eg(z_t, \theta) \approx \frac{1}{T}\sum_{t=1}^{T} g(z_t, \theta) = g_T(\theta)$$

Our estimate, $\hat{\theta}$ will be such that $g_T(\hat{\theta}) \approx 0$. Let's suppose θ is $k \times 1$ and $g_T(\theta)$ is $n \times 1$. If $n < k$, then θ is not identified. If $n = k$ (and $\frac{dg}{d\theta}(\theta_0)$ is full rank), then we are just identified, and we may find $\hat{\theta}$ such that $g_T(\hat{\theta}) = 0$. If $n > k$ (and the rank of $g'(\theta_0) > k$), then we are overidentified. In this case, it will generally be impossible to find $\hat{\theta}$ such that $g_T(\hat{\theta}) = 0$, so we instead minimize a quadratic form

$$\hat{\theta} = \arg\min_{\theta} g_T(\theta)' W_T g_T(\theta)$$

where W_T is symmetric and positive definite.

Some things we want to know:

1. Asymptotics of $\hat{\theta}$
2. Efficient choice of W_t
3. Test of overidentifying restrictions

Assumptions

1. Parameter space is compact
2. $W_T \xrightarrow{p} W$
3. For function $g_0(\theta) = Eg(z_t, \theta)$ assume that $g_0(\theta_0) = 0$ and $Wg_0(\theta) \neq 0$ for any $\theta \neq \theta_0$ (identification)
4. Function $g_0(\theta)$ is continuous
5. Assume that $g_T(\theta)$ converges uniformly in probability to $g_0(\theta)$

Theorem 1. *Under assumptions 1-5 estimator $\hat{\theta}$ is consistent, that is, $\hat{\theta} \xrightarrow{p} \theta_0$.*

If z_t is strictly stationary and ergodic sequence, $g(z_t, \theta)$ is continuous at each θ with probability one and there is $d(z)$ such that $\|g(z, \theta)\| \leq d(z)$ and $Ed(z) < \infty$, then Assumptions 4 and 5 above are satisfied.

Assumptions

6. θ_0 is in the interior of the parameter space and $g_T(\theta)$ is continuously differentiable in the neighborhood of θ_0
7. $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} g(z_t, \theta_0) \Rightarrow N(0, S)$
8. There is $R(\theta)$ continuous at θ_0 and such that $\frac{1}{T} \sum_{t=1}^{T} \frac{\partial g}{\partial \theta}(z_t, \theta) \xrightarrow{p} R(\theta)$ uniformly
9. For $R = R(\theta_0)$ we have that RWR' is non-singular

Note that we're just assuming that some CLT and LLN apply to the above quantities. We can deal with non-iid observations as long as a CLT and LLN apply.

This is the simplest possible setup, for example it requires g to be differentiable, this condition can be relaxed.

Asymptotic Distribution

Theorem 2. *Under assumptions 1-9, $\hat{\theta}$ is asymptotically normal.*

$$\sqrt{T}(\hat{\theta}-\theta) \Rightarrow N(0,\Sigma)$$
$$\Sigma = (RWR')^{-1}(RWSW'R')(RWR')^{-1}$$

Proof. (Sketch of the proof) To prove this, we take a Taylor expansion of the first order condition. The first order condition is:

$$g_T(\theta)'W_T\frac{\partial g_T(\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = 0$$

Expanding $g_T(\theta)$ around θ_0

$$\sqrt{T}g_T(\hat{\theta}) = \sqrt{T}g_T(\theta_0) + \frac{\partial g(\theta_0)}{\partial \theta}\sqrt{T}(\hat{\theta}-\theta_0) + o_p(1)$$

We've assumed that $\sqrt{T}g_T(\theta_0) \Rightarrow N(0,S)$, $\frac{\partial g(\theta_0)}{\partial \theta} \xrightarrow{p} R$. Then, putting our expansion into the first order condition, rearranging, and using these convergence results, we get

$$\begin{aligned}\sqrt{T}(\hat{\theta}-\theta_0) =& (\frac{\partial g_T}{\partial \theta}' W_T \frac{\partial g_T}{\partial \theta})^{-1}(\frac{\partial g_T}{\partial \theta}' W_T \sqrt{T} g_T) \\ \Rightarrow& (RWR')^{-1}RWN(0,S)\end{aligned}$$

□

Efficient Weighting Matrix

Lemma 3. *The efficient choice of W is S^{-1}. This choice of W gives asymptotic variance $\tilde{\Sigma} = (RS^{-1}R)^{-1}$. For any other W, $\Sigma - \tilde{\Sigma}$ is positive semi-definite.*

In practice, we do not know S. There are several options available:

- Feasible efficient GMM:

  Choose arbitrary W, say I, get initial estimate θ_1, use it to calculate $\hat{S}$ (in time series, we need to use Newey-West to estimate $\hat{S}$), then re-run GMM with $W_T = \hat{S}^{-1}$
- We can also iterate this procedure, or put $S(\theta)$ to do *continuous updating estimator* (CUE) GMM.

Under assumptions above these alternatives have the same asymptotic variance. In IV Feasible efficient GMM corresponds to TSLS, CUE is the same as LIML.

Tests of Over-identifying Restrictions When we have more moment conditions than unknown parameters we may test whether model is correctly specified (that is, whether the moments conditions jointly hold).

$$J = [\sqrt{T}g(z,\hat{\theta})]'\hat{\Sigma}^{-1}[\sqrt{T}g(z,\hat{\theta})] \Rightarrow \chi^2_{n-k}$$

Under the null that all moment conditions are true, the statistic, J, converges to a χ^2_{n-k}. Large values of J-test lead to rejection.

Remark 4. This is a test of all the restrictions jointly. If we reject, it does not tell which moment condition is wrong. It merely means the moment conditions contradict one another.

Weak IV

Weak IV problem arises when due to weak correlation between instruments and regressor causes the distributions of IV estimators and test statistics are poorly approximated by their standard asymptotic distributions (normal and chi-squared). The problem is not directly related to non-identification or partial identification (when only part of the unknown parameter is identified). We assume that all the parameters are identified in classical sense. The issue concerns only the quality of classical asymptotics.

Consider a classical setup (with one endogenous regressor, no covariates for simplicity):

$$y_t = \beta x_t + u_t$$
$$x_t = Z_t \pi + v_t,$$

where x_t, y_t are one-dimensional, Z_t is $k \times 1$ and $Eu_t Z_t = 0$.

A GMM estimator for this moment condition solves the following minimization problem: $(y-\beta x)' ZWZ'(y-\beta x) \to \min_\beta$ and thus, $\hat{\beta} = (x'ZWZ'x)^{-1}(x'ZWZ'y)$. The optimal matrix is $W = (Z'Z)^{-1}$. It leads to the estimator called two-stage least squares (TSLS). The usual TSLS will be $\hat{\beta}_{TSLS} = \frac{x'P_Z y}{x'P_Z x}$, where $P_Z = Z(Z'Z)^{-1}Z'$. The problem of weak identification arises when moment conditions are not very informative about the parameter of interest $\frac{\partial g(\theta_0)}{\partial \theta}|_{\theta=\theta_0} = R(\theta_0) = Z'X$ is small (weak correlation between instruments and regressor). Below are some known applied examples where the problem arises (Stock and Watson mini-course at NBER Summer Institute is an excellent reference).

Return to education and quarter of birth. In Angrist Kreuger (1991), y_i is log of earning, x_i is years of education, z_i quarter of birth. Read the paper for explanations. Bounder, Jaeger, Baker (1995) showed that if one randomly assigns quarter of birth and run IV s/he would obtain results close to initial (but instruments are irrelevant in this case!!!). What is important here is that Angrist and Kreuger obtained quite narrow confidence sets (which usually indicates high accuracy of estimation), in a situation when there are little information in the data. It raises suspicions that those confidence sets have poor coverage and are based on poor asymptotic approximations. Nowdays, we can assign this example to weak or many instruments setting.

Elasticity of inter-temporal substitution . Let Δc_{t+1} be consumption growth from t to $t+1$, $r_{i,t+1}$ be a return on asset i from t to $t+1$. The linearized Euler equation is

$$E\left[(\Delta c_{t+1} - \tau_i - \psi r_{i,t+1})|I_t\right] = 0$$

For any variable Z_t measurable with respect I_t we have

$$E\left[(\Delta c_{t+1} - \tau_i - \psi r_{i,t+1})Z_t\right] = 0$$

Our goal is to estimate ψ. It could be done in two ways:

- Run IV (with instruments Z_t) in:
$$\Delta c_{t+1} = \tau_i + \psi r_{i,t+1} + u_t$$
- Run IV (with instruments Z_t) in:
$$r_{i,t+1} = \mu_i + \gamma \Delta c_{t+1} + u_t, \quad \psi = 1/\gamma$$

Finding in literature: 95% confidence sets obtained by two methods do not intersect(!!!) Explanation: the second regression suffers from weak instrument problem, since it is routinely very difficult to predict change in consumption.

Philips curve .

$$\pi_t = \lambda x_t + \gamma_f E_t \pi_{t+1} + \gamma_b \pi_{t-1} + e_t$$

GMM estimated moment condition:

$$E\left[(\pi_t - \lambda x_t - \gamma_f \pi_{t+1} + \gamma_b \pi_{t-1}) Z_{t-1}\right] = 0 \quad Z_{t-1} \in I_{t-1}$$

It is very different to predict π_{t+1} beyond π_{t-1} by any variables observed at time $t-1$, that causes weak relevance of any instruments.