Moment Condition, Identification, and Point Estimation (I):

From Estimating Equation to GMM

Yuan Liao

ORFE

November 17, 2011

Outline

Research Motivation

From Estimating Equation to GMM

Moment Condition and Estimating Equations

Identification: partial, exact, and over

GMM and EL

More on Partial identification

Conclusion

High Dimensional Variable Selection Problem

- Consider a high dimensional variable selection problem:

$$y = x^T \beta_0 + \epsilon, \quad dim(\beta_0) = p >> n$$

$\beta_0 = (\beta_{0S}^T, \beta_{0N}^T)^T$, where $\beta_{0N} = 0$, $\dim(\beta_{0S}) = s << n$.

- Sparse estimation: minimizing an objective function:

$$\hat{\beta} = \arg\min L(\beta) + Penalty(\beta)$$

- Oracle property: $\hat{\beta} = (\hat{\beta}_S^T, \hat{\beta}_N^T)^T$:
  1. $\hat{\beta}_S \to^p \beta_{S0}$,
  2. $\hat{\beta}_N = 0$ with probability approaching one.

Research Motivation

- Usually OLS+penalty is used:

$$\hat{\beta} = \arg\min \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + Penalty$$

The objective function $L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$.

- The literature has been mainly focusing on Penalty(β).

$$\sqrt{n}(\hat{\beta}_S - \beta_{0S}) \to^d N(0, V).$$

V depends on L but not on penalty.

- Can we do something on the objective function as well?

- Soon I noticed that Jelena has done something on $L(\beta)$. Bradic, Fan and Wang (2010) proposed a robust objective function "that is applicable to a large collection of error distribution".
- $\hat{\beta}_{S,BFW} \to^d N(0, V)$. They compared V of $\hat{\beta}_{S,BFW}$ with other methods numerically.
- Is there a best L that we can use, in the sense of minimizing V?

  Semiparametric Efficiency

Efficiency

- Cramer Rao's bound: Suppose $\sqrt{n}(\hat{\beta} - \beta_0) \to^d N(0, V)$, where $\dim(\beta)$ is fixed. Then

$$V \geq I(\beta_0)^{-1},$$

  i.e., up to the leading order, we can do no better than MLE.
- Semiparametric efficiency was introduced by Stein (1956), and was developed by Bickel (1982), Bickel Klaassen, Ritov and Wellner (1990), etc.
- At this point, you can think of it as a bound similar to CR-bound when the likelihood function is not available.

My Research Questions

1. Sufficient conditions for oracle properties:

   For a general objective function $L(\beta)$, under what conditions minimizing $L+$Penalty achieves the oracle?

2. Efficiency:

   For $y = x^T\beta + \epsilon$, but the distribution of ϵ is unknown, what L should be used such that

$$\arg\min L + Penalty$$

   gives the minimum asymptotic variance?

Outline

Moment Condition

- Consider simple linear model WITHOUT variable selection

$$y = x^T \beta_0 + \epsilon, E(\epsilon) = 0$$

  The distribution of ϵ is unknown.

- In addition, assuming: $E(\epsilon x) = 0$, we have

$$E((y - x^T \beta_0)x) = 0 \quad \textbf{Moment Condition}$$

- Estimation: simply replace $E \to \frac{1}{n}\sum_{i=1}^{n}$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T \hat{\beta})x_i = 0(\textbf{ Estimating Equation}).$$

Estimating Equation

- Recall: min $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$: taking derivative:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})x_i = 0 \Rightarrow \text{ OLS} \Leftrightarrow \text{ EE.}$$

- Estimating Equation: Suppose $Em(y, x, \beta_0) = 0$, obtain an estimator by solving:

$$\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \hat{\beta}) = 0.$$

Other Examples of Moment Conditions

- MLE: Under regularity conditions,

$$E\left(\frac{\partial}{\partial \beta} \log L(\beta_0)\right) = 0.$$

$\frac{1}{n}\sum_{i=1}^{n} \partial \log L(\hat{\beta}) = 0 \Rightarrow$ MLE$\Leftrightarrow$ EE.

- GLS: Suppose $y = x^T\beta_0 + \epsilon$, $E(\epsilon|x) = 0$. For any function $f(x)$,

$$E(\epsilon f(x)) = 0 \Leftrightarrow E((y - x^T\beta_0)f(x)) = 0.$$

EE: $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})f(x_i) = 0$. In particular,

$$f(x) = x \Rightarrow \hat{\beta} = OLS.$$

$$f(x) = Var(\epsilon)^{-1}x \Rightarrow \hat{\beta} = GLS: \text{ smaller variance than OLS.}$$

Identification

- Still consider simple linear model:

$$y = x^T\beta_0 + \epsilon, E(\epsilon) = 0$$

$\boxed{x \perp \epsilon,\ E(\epsilon|x) = 0, \text{ or } E(\epsilon x) = 0}$ is important.

- With any of the above three, we have moment condition:

$$E((y - x^T\beta_0)x) = 0$$

Number of Unknowns = Number of equations.

β_0 is uniquely determined (**identified**), i.e.,

$$E((y - x^T\beta)x) = 0 \text{ iff } \beta = \beta_0 = (Exx^T)^{-1}Exy$$

Simply solve $\frac{1}{n}\sum_i (y_i - x_i^T\beta)x_i = 0$.

Three types of Identification

$Em(X, \beta) = 0$

1. Exact Identification: Usually in this case, $\dim(m) = \dim(\beta)$.

   we can solve $\frac{1}{n}\sum_i m(X_i, \hat{\beta}) = 0$. **OLS, MLE, EE**, etc.

2. Partial Identification: $Em(X, \beta) = 0$, but $\dim(m) < \dim(\beta)$:

   More unknowns than equations.

   Solving $\frac{1}{n}\sum_i m(X_i, \hat{\beta}) = 0$ gives infinitely many solutions.

   e.g., $\frac{1}{n}\sum_i y_i - x_i^T \beta = 0$ if $\dim(\beta) > 1$.

   Surprisingly, partial identification exists **almost everywhere** (my Ph.D. thesis). Unfortunately, it has been avoided all the time, partially due to the lack of point estimation consistency.

3. Over Identification: **GMM/ EL**.

Over Identification

- Refers to the case: β_0 is uniquely determined by $Em(X, \beta_0) = 0$, but usually $\dim(m) > \dim(\beta)$.

  No solution for: (more equations than unknowns)

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \beta) = 0.$$

  Therefore, EE does not work.

- Examples of over-identification:

  $y = x^T\beta_0 + \epsilon$, $E(\epsilon|x) = 0$. We have:

$$E((y - x^T\beta_0)x_i^k) = 0, i = 1, ..., p, \text{ and } k = 1, 2, ...$$

Instrumental Variables

Wage regression in labor economics:

$$\log(wage) = \beta_0 + \beta_1(\text{ years of education}) + \epsilon.$$

$y = \beta_0 + \beta_1 x + \epsilon$. An overview: Card (1999).

- Other variables are also correlated with wage, e.g., family wealth. Thus the above equation with the assumption $E(\epsilon|x) = 0$ is a mis-specified model.
- When $E(x\epsilon) \neq 0$, x is **endogenous**; o.w. is **exogenous**.
- The biggest problem is the lack of identification.

Instrumental Variables

$y = x^T\beta_0 + \epsilon, E(x\epsilon) \neq 0.$

- We observe $w = (w_1, ..., w_k)$. $E(\epsilon w) = 0$.
- $E((y - x^T\beta_0)w) = 0$. If $\dim(w) \geq \dim(\beta_0)$, β_0 is identified. (Rigorously, $rank(Ewx^T) = \dim(\beta_0)$.)
- w is called Instrumental Variable (IV).
- If β_0 is identified and $\dim(w) > \dim(\beta_0)$: over-identification.
- Of course, we can discard some IV's so that $\dim(w) = \dim(\beta_0)$: exact identification.
- However, we lose some information$\Rightarrow$ large variance.

Generalized Method of Moments

- Suppose β_0 is uniquely determined by $Em(X, \beta_0) = 0$

  Equivalently, for positive definite W,

$$Em(X, \beta)^T WEm(X, \beta) = 0 \text{ iff } \beta = \beta_0,$$

  β_0 is the unique minimizer of Q.

- Hansen (1982 *Econometrica*):

$$\hat{\beta}_{GMM} \equiv \arg\min \frac{1}{n}\sum_{i=1}^{n} m(X_i, \beta)^T W \frac{1}{n}\sum_{i=1}^{n} m(X_i, \beta).$$

  Note that GMM allows $\dim(m) > \dim(\beta)$, but EE does not. But when $\dim(m) \leq \dim(\beta)$, GMM=EE.

Example

$$y = x^T \beta_0 + \epsilon, E(\epsilon) = 0.$$

Suppose $E(\epsilon|x) = 0$:

- For any $k \times 1$ vector function $f(x)$, $k \geq \dim(\beta_0)$,

$$\hat{\beta}_{GMM} = \arg\min \frac{1}{n} \sum_{i=1}^{n} [(y_i - x_i^T \beta) f(x_i)]^T W \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^T \beta) f(x_i).$$

- If $f(x) = x$, GMM=OLS.
- Good choice of W and $f(x)$ yields smaller variance than OLS.
- Each choice $(f, W) \Rightarrow \sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \to^d N(0, V)$.
- There exist best (f^*, W^*), depending on $Var(\epsilon|x)$.

Example

$$y = x^T\beta_0 + \epsilon, E(\epsilon) = 0.$$

Suppose $E(\epsilon|x) \neq 0$:

- $E(\epsilon|x) \neq 0$, but $E(\epsilon w) = 0$, $\dim(w) \geq \dim(\beta_0)$

  Famous example: $y = \log(wage)$, $x = Edu$, $w =$ distance.

$$\hat{\beta}_{GMM} = \arg\min \frac{1}{n}\sum_{i=1}^{n}[(y_i - x_i^T\beta)w_i]^T W \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)w_i.$$

- W^* : minimizes asymptotic variance.
- When $W = W^*$, $\hat{\beta}_{GMM}$ is equivalent to **two stage least square**.

2SLS

$$y = x^T\beta_0 + \epsilon,$$

$E(\epsilon|x) \neq 0$, but $E(\epsilon w) = 0$.

- We can write $u = x - \pi$, where π is such that $Eu = 0$.
- Assume $E(uw) = 0$:

$$y = x^T\beta_0 + \epsilon, \quad x = \pi w + u.$$

- **Stage 1** OLS$\Rightarrow \hat{\pi} \Rightarrow \hat{x} = \hat{\pi} w$

  Stage 2 OLS on $(y, \hat{x}) \Rightarrow \hat{\beta}_{2SLS}$.

Empirical Likelihood

1. EL is an alternative to GMM.
2. Suppose β_0 is identified by $Em(X, \beta_0) = 0$:
3. Owen (1990 *Annals*):

$$L(\beta) = \max \prod_{i=1}^{n} p_i$$

$$s.t. p_i \geq 0, \sum_{i=1}^{n} p_i = 1, \sum_{i=1}^{n} p_i m(X_i, \beta) = 0.$$

$$\hat{\beta}_{EL} = \arg\max L(\beta).$$

Large Sample Properties

- Under regularity conditions, we have consistency and asym. norm. for both GMM and EL.
- In particular,

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \to^d N(0, V(W)).$$

Let $W^* = \arg\min V(W)$, then

$$\sqrt{n}(\hat{\beta}_{EL} - \beta_0) \to^d N(0, V(W^*)).$$

- In general, more moment conditions$\Rightarrow$ smaller variance.

Semiparametric Efficiency

- Suppose $Em(X, \beta_0) = 0$. Let $W^* = \arg\min V(W)$. Chamberlain (1987 *Journal of Econometrics*):

$$\arg\min \frac{1}{n}\sum_{i=1}^{n} m(X_i, \beta)^T W^* \frac{1}{n}\sum_{i=1}^{n} m(X_i, \beta)$$

  is the best thing we can do, if only $Em(X, \beta_0) = 0$ is known, instead of the likelihood.

- So is $\hat{\beta}_{EL}$.

Outline

More on Partial identification

- As we've seen, if $Em(X, \beta_0) = 0$, and $\dim(m) < \dim(\beta)$, β_0 is not identified, e.g., $y = x^T\beta_0 + \epsilon$, $E(\epsilon) = 0$.
- One important example: moment inequality:

$$Em(X, \beta_0) \geq 0.$$

Equivalently, $Em(X, \beta_0) - \lambda = 0$ for $\lambda \geq 0$.

$$Eg(X, \beta_0, \lambda) = 0.$$

- Partial identification is the most robust model.
- However, there is no consistent point estimation.

More Examples of Moment Inequality

Missing Data & Causal Effect $y \in \{0, 1\}$, but subject to missing.

We want to estimate $\beta = P(y = 1)$.

$$P(y = 1) = P(y = 1|missing)P(missing)$$

$$+P(y = 1|notmissing)P(notmissing).$$

Missing at random: $P(y = 1) = P(y = 1|notmissing)$

More robust: $P(y = 1|notmissing)P(notmissing) \leq \beta$

$\leq P(y = 1|notmissing)P(notmissing) + P(missing)$.

Censored Data $y = x^T\beta_0 + \epsilon$, observe $Z = \min\{y, C\}$.

Assume $Median(\epsilon|x) = 0$. Khan and Tamer (2009, *JOE*):

$$
\begin{aligned}
E(I(Z \geq x^T\beta_0)|x) &= P(Z \geq x^T\beta_0|x) = P(y \geq x^T\beta_0, C \geq x^T\beta_0|x) \\
&= P(\epsilon \geq 0, C \geq x^T\beta_0|x) \leq P(\epsilon \geq 0|x) \\
&= 0.5
\end{aligned}
$$

Let $m(X,\beta_0) = I(Z \geq x^T\beta_0) - 0.5 \Rightarrow E(m(X,\beta)|x) \geq 0$.

For any $f(x) \geq 0$, $Em(X,\beta_0)f(x) \geq 0$.

English Auction $y = x^T\beta_0 + \epsilon$, $E(\epsilon|x) = 0$.

y : valuation: max value a bidder is willing to pay, unobservable.

x : object, organization, income..., observable.

(y_1, y_2): (bidder's final bid, winning bid): observable.

Δ : minimum increment: observable.

It is known $y_1 \leq y \leq y_2 + \Delta$.

$$E(y_1|x) \leq E(y|x) = x^T\beta_0 \leq E(y_2 + \Delta|x).$$

Is β_0 identified? How do we estimate it/ inference? (known)

Variable selection? (unknown)

Conclusion

- Three types of identification:
  1. exact: regular simple linear model, nonlinear model
  2. over: EE does not work
  3. partial: most robust. Moment inequality
- Need to be careful when assuming either $E(x\epsilon) = 0$ or $E(\epsilon|x) = 0$.
- GMM and EL
  - Semiparametric efficient
  - Can be used as alternative objective functions to LS.