Set-up

Let $m(x, \theta)$ be a parametric model for $E(y|x)$ (we will also use it later for other conditional moments, such as $Q_\tau(y|x)$ in quantile regression) where

- $m$ is a known function of $x$ and $\theta$.
- $x$ is a $K$ vector of explanatory variables.
- $\theta$ is a $P \times 1$ parameter vector. $\theta \in \Theta \subseteq \mathbb{R}^P$.

A model is **correctly specified** for the conditional mean $E(y|x)$, if for some $\theta_0 \in \Theta$,

$$E(y|x) = m(x, \theta_0)$$
Set-up

- Let $q(w, \theta)$ be a function of the random vector $w$ and the parameter vector $\theta \in \Theta$.
  - In general we will have $w = (y, x)$ with typical element $w_i$ from a sample $\{w_i : i = 1, 2, \ldots, N\}$.
  - $q$ is a known function of $w$ and $\theta$.
  - $\theta$ is a $P \times 1$ parameter vector. $\theta \in \Theta \subseteq \mathbb{R}^P$.
- An M-estimator of $\theta_0$ solves the problem
  \[
  \min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(w_i, \theta),
  \]
  
  \[
  \hat{\theta} = \arg \min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(w_i, \theta)
  \]
- Define the function $\theta \mapsto M_N(\theta)$ where $M_N(\theta) \equiv N^{-1} \sum_{i=1}^{N} q(w_i, \theta)$ and $\theta \mapsto M(\theta)$ where $M(\theta) \equiv E[q(w, \theta)]$.
- Note that $\hat{\theta}$ is a random variable as it depends on the sample $\{w_i : i = 1, 2, \ldots, N\}$. Sometimes we will use the notation $\hat{\theta}_N$ to emphasize that it depends on the sample.
It is generally assumed that

$$\theta_0 = \arg \min_{\theta \in \Theta} E[q(w, \theta)].$$

$E[q(w, \theta)]$ can be seen in statistical terms a loss function.

Examples of loss functions are:
- quadratic: $(.)^2$ which is the base of OLS estimators;
- absolute value: $|.|$ which is the base of median regression estimators.
Identification requires that there is a unique solution to the minimisation problem:

\[ E[q(w, \theta_0)] < E[q(w, \theta)], \text{ for all } \theta \in \Theta, \theta \neq \theta_0. \]
Because $\hat{\theta}$ depends on the whole function $\theta \mapsto M_N(\theta)$ we require an appropriate "functional convergence".

We require uniform convergence in probability:

$$\max_{\theta \in \Theta} \left| N^{-1} \sum_{i=1}^{N} q(w_i, \theta) - E[q(w, \theta)] \right| \overset{p}{\rightarrow} 0.$$  

Uniform convergence is a difficult and technical issue.
Uniform Weak Law of Large Numbers: Let \( w \) be a random vector taking values in \( \mathcal{W} \subset \mathbb{R}^M, \Theta \subset \mathbb{R}^P \), and let \( q : \mathcal{W} \times \Theta \to \mathbb{R} \) be a real valued function. Assume that
(a) \( \Theta \) is compact;
(b) for each \( \theta \in \Theta \), \( q(., \theta) \) is Borel measurable on \( \mathcal{W} \);
(c) for each \( w \in \mathcal{W} \), \( q(w, .) \) is continuous on \( \Theta \); and
(d) \( |q(w, \theta)| < b(w) \) for all \( \theta \in \Theta \), where \( b \) is a nonnegative function on \( \mathcal{W} \) such that \( E[b(w)] < \infty \).

Then
\[
\max_{\theta \in \Theta} \left| N^{-1} \sum_{i=1}^{N} q(w_i, \theta) - E[q(w, \theta)] \right| \xrightarrow{P} 0.
\]
Consistency of M-estimators: Under the assumptions required for uniform convergence in probability above, and assuming identification holds, then a random vector $\hat{\theta}$ that is an M-estimator (i.e. $\hat{\theta} = \arg\min_{\theta \in \Theta} N^{-1} \sum_{i=1}^{N} q(w_i, \theta)$) satisfies $\hat{\theta} \xrightarrow{p} \theta$.

Suppose that $\hat{\theta} \xrightarrow{p} \theta$, and assume that $r(w, \theta)$ satisfies the same assumptions on $q(w, \theta)$. Then, $N^{-1} \sum_{i=1}^{N} r(w, \hat{\theta}) \xrightarrow{p} E[r(w, \theta_0)]$. 
Example: OLS

- In OLS models \( m(x, \beta) = E(y|x) = x\beta \) and \( q(y, x, \beta) = [y - m(x, \beta)]^2 \).
- For identification of OLS models we solve

\[
E[q(y, x, \beta_0)] = E[y^2 + m(x, \beta_0)^2 - 2ym(x, \beta_0)^2] \\
= E[(x\beta_0)^2 + u^2 + (x\beta_0)^2 - 2(x\beta_0)^2] = \sigma^2
\]

\[
E[q(y, x, \beta)]_{\beta \neq \beta_0} = E[y^2 + m(x, \beta)^2 - 2ym(x, \beta)^2] \\
= E[(x\beta_0)^2 + u^2 + (x\beta)^2 - 2(x\beta_0)(x\beta)] \\
= \sigma^2 + E(x\beta_0 - x\beta)^2
\]

Then,

\[
E[q(x, \beta_0)] < E[q(x, \beta)] \text{ for all } \beta \in \mathbf{B}, \beta \neq \beta_0.
\]
Score of the objective function

If \( q(w, .) \) is continuously differentiable on the interior of \( \Theta \), then with probability approaching one, \( \hat{\theta} \) solves the first-order condition

\[
\sum_{i=1}^{N} s(w_i, \hat{\theta}) = 0,
\]

where

\[
s(w, \theta)' = \nabla_\theta q(w, \theta) = \left( \frac{\partial q(w, \theta)}{\partial \theta_1}, \frac{\partial q(w, \theta)}{\partial \theta_2}, \ldots, \frac{\partial q(w, \theta)}{\partial \theta_P} \right)'
\]

is the score of the objective function. The condition above is usually satisfied with equality (but not always, see OLS vs. quantile regression).
Score of the objective function

- If $E[q(w, \theta)]$ is continuously differentiable on the interior of $\Theta$, then a condition for the maximisation is:
  \[
  \nabla_\theta E[q(w, \theta)]_{\theta=\theta_0} = 0.
  \]

- If the derivative and expectations can be interchanged (Dominated Convergence Theorem), then
  \[
  E[\nabla_\theta q(w, \theta_0)] = E[s(w, \theta_0)] = 0.
  \]

- Moreover, if $\hat{\theta} \xrightarrow{P} \theta$, then under standard regularity conditions
  \[
  N^{-1} \sum_{i=1}^{N} s(w_i, \hat{\theta}) \xrightarrow{P} E[s(w, \theta_0)].
  \]

An estimator that is defined by

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} \| \sum_{i=1}^{N} s(w_i, \theta) \|_M
\]

is called a **Z-estimator** (for zero). Here $\| \|_M$ is a specified norm (see GMM). For this case we should assume a different identification condition

\[
E[s(w, \theta)] = 0 \text{ iff } \theta = \theta_0.
\]
Hessian of the objective function

- If \( q(w, \cdot) \) is twice continuously differentiable on the interior of \( \Theta \), then define the Hessian of the objective function,

\[
H(w, \theta) = \nabla^2 q(w, \theta) = \frac{\partial^2 q(w, \theta)}{\partial \theta \partial \theta'}
\]

- By the mean-value expansion of \( q(w, \theta) \) expanded about \( \theta_0 \),

\[
\sum_{i=1}^{N} s(w_i, \hat{\theta}) = \sum_{i=1}^{N} s(w_i, \theta_0) + \left( \sum_{i=1}^{N} \hat{H}_i \right) (\hat{\theta} - \theta_0),
\]

where \( \hat{H}_i \equiv H(w_i, \tilde{\theta}) \) and \( \tilde{\theta} \) is a vector where each element is on the line segment between \( \hat{\theta} \) and \( \theta_0 \).

- Using similar arguments as for the score if \( \hat{\theta} \xrightarrow{p} \theta \), then by asymptotic theory

\[
N^{-1} \sum_{i=1}^{N} H(w_i, \hat{\theta}) \xrightarrow{p} E[H(w, \theta_0)].
\]

- For M-estimators, if \( \theta_0 \) is identified then \( E[H(w, \theta_0)] \) is positive definite.
The expansion about the true value $\theta_0$ plays a fundamental role in asymptotic analysis. Rearranging terms we obtain the first-order asymptotic expansion of $\hat{\theta}$:

$$\sqrt{N}(\hat{\theta} - \theta_0) = \left( N^{-1} \sum_{i=1}^{N} \hat{H}_i \right)^{-1} \left[ -N^{-1/2} \sum_{i=1}^{N} s_i(\theta_0) \right]$$

where $s_i(\theta_0) = s(w_i, \theta_0)$.

Moreover, using $A_0 \equiv E[H(w, \theta_0)]$,

$$\sqrt{N}(\hat{\theta} - \theta_0) = A_0^{-1} \left[ -N^{-1/2} \sum_{i=1}^{N} s_i(\theta_0) \right]$$

$$= A_0^{-1} \left[ -N^{-1/2} \sum_{i=1}^{N} s_i(\theta_0) \right] + N^{-1/2} \sum_{i=1}^{N} s_i(\theta_0) + o_p(1) \cdot O_p(1)$$

Define the influence function of $\hat{\theta}$ as $e(w_i, \theta_0) \equiv e_i(\theta_0) \equiv A_0^{-1}s_i(\theta_0)$ because this provides the “influence” of each particular $i-th$ observation on the estimator.
Asymptotic normality: In addition to the assumptions for consistency assume that
(a) \( \theta_0 \in \text{int}\Theta \);
(b) \( s(w,.) \) is continuously differentiable on the interior of \( \Theta \) for all \( w \in \mathcal{W} \);
(c) Each element in \( H(w, \theta) \) is bounded in absolute value by a function \( b(w) \), where \( E[b(w)] < \infty \);
(d) \( A_0 \equiv E[H(w, \theta_0)] \) is positive definite;
(e) \( E[s(w, \theta_0)] = 0 \); and
(f) Each element in \( s(w, \theta_0) \) has finite second moment.

Then,

\[
\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, A_0^{-1}B_0A_0^{-1}),
\]

where \( B_0 \equiv E[s(w, \theta_0)s(w, \theta_0)'] = \text{Var}[s(w, \theta_0)] \). Thus,

\[
\text{Avar}(\hat{\theta}) = A_0^{-1}B_0A_0^{-1} / N.
\]

Note the sandwich formula: \( A_0^{-1}B_0A_0^{-1} \). This will appear many times later on.
The asymptotic expansion above can also be used for functions of parameters. Consider a function \( \theta \mapsto g(\theta) \). We would like to know what is the distribution of \( \sqrt{N}(g(\hat{\theta}) - g(\theta_0)) \).

Assume that
- \( \sqrt{N}(\hat{\theta} - \theta_0) \overset{d}{\to} N(0, \hat{V}_\theta) \),
- \( \theta \mapsto g(\theta) \) is differentiable at \( \theta_0 \) (this means the first derivatives exist and are continuous at \( \theta_0 \)).

Then,

\[
\sqrt{N}(g(\hat{\theta}) - g(\theta_0)) \overset{d}{\to} N(0, \nabla_{\theta} g(\theta_0) \hat{V}_\theta \nabla_{\theta} g(\theta_0)')
\]

Note that by Slutsky’s theorem:

\[
\nabla_{\theta} g(\hat{\theta}) \hat{V}_\theta \nabla_{\theta} g(\hat{\theta})' \overset{p}{\to} \nabla_{\theta} g(\theta_0) \hat{V}_\theta \nabla_{\theta} g(\theta_0)'
\]

where \( \hat{V}_\theta \) is a consistent estimator of \( V_\theta \).
Example: OLS

- Note that for OLS models

\[ s(w, \beta) = \nabla_\theta q(w, \beta) = -2x'(y - x\beta), \]

which gives us the classical conditions \( E[x'u] = 0. \)

- The Hessian in this case is:

\[ A_0 = E[H(w, \beta_0)] = 2E[x'x]. \]

- Moreover, assuming homoskedasticity,

\[ B_0 = E[s(w, \beta_0)s(w, \beta_0)'] = 4\sigma^2 E[x'x] \]

- Then,

\[ \sqrt{N}(\hat{\beta} - \beta_0) \overset{d}{\rightarrow} N(0, \sigma^2 E[x'x]^{-1}). \]
Example: OLS

- If there is heteroskedasticity:

  \[ B_0 = E[s(w, \beta_0) s(w, \beta_0)'] = 4E[x' u' ux] \]

  and then \( B_0 \neq A_0 \).

- For this case,

  \[ \sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, E[x'x]^{-1} E[x' u' ux] E[x'x]^{-1}) \].
Consider a regression model $y_t = \beta_0 + x_t \beta_1 + y_{t-1} \alpha_1 + u_t$ where $(y, x)$ are in logarithm.

- The short-run effect (elasticity) of $x$ on $y$ is given by $\beta_1$.
- The long-run effect (elasticity) of $x$ on $y$ is given by $\frac{\hat{\beta}_1}{1 - \hat{\alpha}_1}$.
- Obtain the distribution of $\frac{\hat{\beta}_1}{1 - \hat{\alpha}_1}$.
Generalized method of moments

- Let \( \mathbf{g}(\mathbf{w}, \theta) \) be a \( L \times 1 \) vector map \( \mathcal{W} \times \Theta \mapsto \mathbf{g}(\mathbf{w}, \theta) \). Let \( \mathbf{g}_i(\theta) \equiv \mathbf{g}(\mathbf{w}_i, \theta) \).
- Assume that \( E[\mathbf{g}(\mathbf{w}_i, \theta_0)] = 0 \).
- A minimal condition to identify \( \theta_0 \) is \( L \geq P \).
- When \( L = P \), then \( \theta_0 \) is estimated by setting the sample counterpart \( N^{-1} \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}, \theta_0) \) to zero. This is the OLS case where \( \mathbf{g}(\mathbf{w}, \theta) = \mathbf{x}'(\mathbf{y} - \mathbf{x}\beta) \).
- When \( L > P \), then the generalized method of moments (GMM) uses an appropriate metric to minimize a quadratic form of \( \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}, \theta_0) \):

\[
\min_{\theta \in \Theta} \left[ \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \theta) \right]' \hat{\Xi} \left[ \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \theta) \right]
\]

where \( \hat{\Xi} \) is an \( L \times L \) symmetric weighting positive semidefinite matrix. This is the case of IV estimators.
- Let \( Q_N(\theta) = \left[ \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \theta) \right]' \hat{\Xi} \left[ \sum_{i=1}^{N} \mathbf{g}(\mathbf{w}_i, \theta) \right] \). Under standard regularity conditions \( Q_N(\theta) \) converges uniformly to \( \{E[\mathbf{g}_i(\theta)]\}' \Xi_0 \{E[\mathbf{g}_i(\theta)]\} \) where \( \hat{\Xi} \xrightarrow{P} \Xi_0 \).
Generalized method of moments

- Under the assumption that $g(w, .)$ is continuously differentiable on $\text{int}(\Theta)$, $\theta_0 \in \Theta$, then the first order condition for $\hat{\theta}$ is

$$\left[ \sum_{i=1}^{N} \nabla_{\theta} g_i(\hat{\theta}) \right] \hat{\Xi} \left[ \sum_{i=1}^{N} g_i(\hat{\theta}) \right] \equiv 0.$$

- Define
  - $L \times P$ matrix with rank $P$, $G_0 \equiv E[\nabla_{\theta} g_i(\hat{\theta})]$;
  - $P \times P$ matrix with rank $P$, $A_0 \equiv G_0' \Xi G_0$;
  - $P \times P$ matrix with rank $P$, $B_0 \equiv G_0' \Xi \Lambda_0 \Xi G_0$;
  - $L \times L$ matrix with rank $L$, $\Lambda_0 \equiv E[g_i(\theta_0)g_i(\theta_0)'] = \text{Var}[g_i(\theta_0)].$

- Then, after some calculations

$$0 = G_0' \Xi_0 N^{-1/2} \sum_{i=1}^{N} g_i(\theta_0) + A_0 \sqrt{N}(\hat{\theta} - \theta_0) + o_p(1)$$

- Then,

$$\sqrt{N}(\hat{\theta} - \theta_0) = -A_0^{-1} G_0' \Xi_0 N^{-1/2} \sum_{i=1}^{N} g_i(\theta_0) + o_p(1) \overset{d}{\rightarrow} N(0, A_0^{-1} B_0 A_0^{-1})$$
Optimal efficient GMM estimator requires \( \Xi_0 = \Lambda_0^{-1} \) because

\[
(G_0' \Xi_0 G_0)^{-1} (G_0' \Xi_0 \Lambda_0 \Xi_0 G_0) (G_0' \Xi_0 G_0)^{-1} - (G_0' \Lambda_0 G_0)^{-1}
\]

i.e. the difference between two hypothesical GMM estimators, can be shown to be positive semidefinite.

Note however that we cannot have an estimator of \( \Lambda_0 \) before we estimate \( \hat{\theta} \). However, we only require any consistent estimator, say \( \tilde{\theta} \) to estimate first \( \Lambda_0 \), then plug in the GMM framework.
References

This slides are based on

- Chapters 12, 13 and 14 of Wooldridge.