Econometrics 2

Generalized Method of Moments (GMM) Estimation

Heino Bohn Nielsen
Outline

(1) Introduction and Motivation

(2) Moment Conditions and Identification

(3) A Model Class: Instrumental Variables (IV) Estimation

(4) Method of Moment (MM) Estimation
   Examples: Mean, OLS and Linear IV

(5) Generalized Method of Moment (GMM) Estimation
   Properties: Consistency and Asymptotic Distribution

(6) Efficient GMM
   Examples: Two-Stage Least Squares

(7) Comparison with Maximum Likelihood
   Pseudo-ML Estimation

(8) Empirical Example: C-CAPM Model
Introduction

Generalized method of moments (GMM) is a general estimation principle. Estimators are derived from so-called moment conditions.

Three main motivations:
(1) Many estimators can be seen as special cases of GMM. Unifying framework for comparison.

(2) Maximum likelihood estimators have the smallest variance in the class of consistent and asymptotically normal estimators. But: We need a full description of the DGP and correct specification. GMM is an alternative based on minimal assumptions.

(3) GMM estimation is often possible where a likelihood analysis is extremely difficult. We only need a partial specification of the model. Models for rational expectations.
Moment Conditions and Identification

- A moment condition is a statement involving the data and the parameters:

\[ g(\theta_0) = E[f(w_t, z_t, \theta_0)] = 0. \]  

where \( \theta \) is a \( K \times 1 \) vector of parameters with true value \( \theta_0 \); \( f(\cdot) \) is an \( R \times 1 \) vector of (non-linear) functions; \( w_t \) contains model variables; and \( z_t \) contains instruments.

- If we knew the expectation then we could solve the equations in (*) to find \( \theta_0 \).

- If there is a unique solution, so that

\[ E[f(w_t, z_t, \theta)] = 0 \quad \text{if and only if} \quad \theta = \theta_0, \]

then we say that the system is identified.

- Identification is essential for doing econometrics. Two ideas:
  (1) Is the model constructed so that \( \theta_0 \) is unique (identification).
  (2) Are the data informative enough to determine \( \theta_0 \) (empirical identification).
**Instrumental Variables Estimation**

- In many applications, the moment condition has the specific form:

\[
f(w_t, z_t, \theta) = \underbrace{u(w_t, \theta)}_{(1 \times 1)} \cdot \underbrace{z_t}_{(R \times 1)},
\]

where the \( R \) instruments in \( z_t \) are multiplied by the disturbance term, \( u(w_t, \theta) \).

- You can think of \( u(w_t, \theta) \) as the equivalent of an error term. The moment condition becomes

\[
g(\theta_0) = E[u(w_t, \theta_0) \cdot z_t] = 0,
\]

stating that the instruments are uncorrelated with the error term of the model.

- This class of estimators is referred to as **instrumental variables estimators**. The function \( u(w_t, \theta) \) may be linear or non-linear in \( \theta \).
Consider a monetary policy rule, where the interest rate depends on expected future inflation:

\[ r_t = \beta \cdot E[\pi_{t+1} \mid I_t]. \]

Noting that

\[ \pi_{t+1} = E[\pi_{t+1} \mid I_t] - v_t, \]

where \( v_t \) is the expectation error, we can write the model as

\[ r_t = \beta \cdot E[\pi_{t+1} \mid I_t] = \beta \cdot \pi_{t+1} + \beta \cdot v_t = \beta \cdot \pi_{t+1} + u_t. \]

Note that \( \pi_{t+1} \) and \( u_t \) are correlated: OLS is inconsistent.

Under rational expectations, the expectation error, \( v_t \), should be orthogonal to the information set, \( I_t \), and for \( z_t \in I_t \) we have the moment condition

\[ E[u_t \cdot z_t] = E[(r_t - \beta \cdot \pi_{t+1}) \cdot z_t] = 0. \]

This is enough to identify \( \beta \).
Method of Moments (MM) Estimator

- For a given sample, \( w_t \) and \( z_t \) \((t = 1, 2, ..., T)\), we cannot calculate the expectation. We replace with sample averages to obtain the analogous sample moments:

\[
g_T(\theta) = \frac{1}{T} \sum_{t=1}^{T} f(w_t, z_t, \theta).
\]

We can derive an estimator, \( \hat{\theta}_{MM} \), as the solution to \( g_T(\hat{\theta}_{MM}) = 0 \).

- To find a unique estimator, we need at least as many equations as parameters. The order condition for identification is \( R \geq K \).
  
  - \( R = K \) is called **exact identification**.
    The estimator is denoted the method of moments estimator, \( \hat{\theta}_{MM} \).
  
  - \( R > K \) is called **over-identification**.
    The estimator is denoted the generalized method of moments estimator, \( \hat{\theta}_{GMM} \).
Example: MM Estimator of the Mean

• Assume that $y_t$ is random variable drawn from a population with expectation $\mu_0$. We have a single moment condition:

$$g(\mu_0) = E[f(y_t, \mu_0)] = E[y_t - \mu] = 0,$$

where $f(y_t, \mu_0) = y_t - \mu_0$.

• For a sample, $y_1, y_2, \ldots, y_T$, we state the corresponding sample moment condition:

$$g_T(\hat{\mu}) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{\mu}) = 0.$$

The MM estimator of the mean $\mu_0$ is the solution, i.e.

$$\hat{\mu}_{MM} = \frac{1}{T} \sum_{t=1}^{T} y_t,$$

which is the sample average.
Example: OLS as a MM Estimator

- Consider the linear regression model of $y_t$ on $x_t$ ($K \times 1$):

$$y_t = x_t'\beta_0 + \epsilon_t.$$  \hfill (**)  

Assume that (**) represents the conditional expectation:

$$E[y_t \mid x_t] = x_t'\beta_0 \quad \text{so that} \quad E[\epsilon_t \mid x_t] = 0.$$  

- That implies the $K$ unconditional moment conditions

$$g(\beta_0) = E[x_t\epsilon_t] = E[x_t (y_t - x_t'\beta_0)] = 0,$$

which we recognize as the minimal assumption for consistency of the OLS estimator.
• We define the corresponding sample moment conditions as

\[ g_T(\hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} x_t \left( y_t - x_t'\hat{\beta} \right) = \frac{1}{T} \sum_{t=1}^{T} x_t y_t - \frac{1}{T} \sum_{t=1}^{T} x_t x_t' \hat{\beta} = 0. \]

And the MM estimator is derived as the unique solution:

\[ \hat{\beta}_{MM} = \left( \frac{1}{T} \sum_{t=1}^{T} x_t x_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} x_t y_t, \]

provided that \( \frac{1}{T} \sum_{t=1}^{T} x_t x_t' \) is non-singular.

• Method of moments is one way to motivate the OLS estimator. Highlights the minimal (or identifying) assumptions for OLS.
Example: Under-Identification

- Consider again a regression model

\[ y_t = x_t' \beta_0 + \epsilon_t = x_{1t}' \gamma_0 + x_{2t}' \delta_0 + \epsilon_t. \]

- Assume that the $K_1$ variables in $x_{1t}$ are predetermined, while the $K_2 = K - K_1$ variables in $x_{2t}$ are endogenous. That implies

\[
E[x_{1t} \epsilon_t] = 0 \ (K_1 \times 1) \tag{†}
\]

\[
E[x_{2t} \epsilon_t] \neq 0 \ (K_2 \times 1). \tag{††}
\]

- We have $K$ parameters in $\beta_0 = (\gamma_0', \delta_0')'$, but only $K_1 < K$ moment conditions (i.e. $K_1$ equations to determine $K$ unknowns). The parameters are not identified and cannot be estimated consistently.
Example: Simple IV Estimator

- Assume $K_2$ new variables, $z_{2t}$, that are correlated with $x_{2t}$ but uncorrelated with $\epsilon_t$:

$$E[z_{2t}\epsilon_t] = 0.$$  \hfill (†††)

The $K_2$ moment conditions in (†††) can replace (††). To simplify notation, we define

$$x_t = \begin{pmatrix} x_{1t} \\ x_{2t} \end{pmatrix}_{(K \times 1)} \quad \text{and} \quad z_t = \begin{pmatrix} x_{1t} \\ z_{2t} \end{pmatrix}_{(K \times 1)}.$$

$x_t$ are model variables, $z_{2t}$ are new instruments, and $z_t$ are instruments. We say that $x_{1t}$ are instruments for themselves.

- Using (†) and (†††) we have $K$ moment conditions:

$$g(\beta_0) = \begin{pmatrix} E[x_{1t}\epsilon_t] \\ E[z_{2t}\epsilon_t] \end{pmatrix} = E[z_t \epsilon_t] = E[z_t (y_t - x_t'\beta_0)] = 0,$$

which are sufficient to identify the $K$ parameters in $\beta$. 
• The corresponding **sample moment conditions** are given by

\[
g_T(\hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} z_t \left( y_t - x_t'\hat{\beta} \right) = 0.
\]

• The method of moments estimator is the unique solution:

\[
\hat{\beta}_{MM} = \left( \frac{1}{T} \sum_{t=1}^{T} z_t x_t' \right)^{-1} \frac{1}{T} \sum_{t=1}^{T} z_t y_t,
\]

provided that \( \frac{1}{T} \sum_{t=1}^{T} z_t x_t' \) is non-singular.

• Note the following:
  (1) We need the instruments to identify the parameters.
  (2) The MM estimator coincides with the simple IV estimator.
  (3) The procedure only works with \( K_2 \) new instruments (i.e. \( R = K \)).
  (4) Non-singularity of \( \sum_{t=1}^{T} z_t x_t' \) requires relevant instruments.
Generalized Method of Moments Estimation

- The case $R > K$ is called over-identification. More equations than parameters and no solution to $g_T(\theta) = 0$ in general.

- Instead we minimize the distance from $g_T(\theta)$ to zero. The distance is measured by the quadratic form

$$Q_T(\theta) = g_T(\theta)' W_T g_T(\theta),$$

where $W_T$ is an $R \times R$ symmetric and positive definite weight matrix.

- The GMM estimator depends on the weight matrix:

$$\hat{\theta}_{GMM}(W_T) = \arg \min_{\theta} \{g_T(\theta)' W_T g_T(\theta)\}.$$
Distances and Weight Matrices

- Consider a simple example with 2 moment conditions

\[ g_T(\theta) = \begin{pmatrix} g_a \\ g_b \end{pmatrix}, \]

where the dependence of \( T \) and \( \theta \) is suppressed.

- First consider a simple weight matrix, \( W_T = I_2 \):

\[ Q_T(\theta) = g_T(\theta)'W_Tg_T(\theta) = \begin{pmatrix} g_a & g_b \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g_a \\ g_b \end{pmatrix} = g_a^2 + g_b^2, \]

which is the square of the simple distance from \( g_T(\theta) \) to zero. Here the coordinates are equally important.

- Alternatively, look at a different weight matrix:

\[ Q_T(\theta) = g_T(\theta)'W_Tg_T(\theta) = \begin{pmatrix} g_a & g_b \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g_a \\ g_b \end{pmatrix} = 2 \cdot g_a^2 + g_b^2, \]

which attaches more weight to the first coordinate in the distance.
Consistency: Why Does it Work?

• Assume that a law of large numbers (LLN) applies to $f(w_t, z_t, \theta)$, i.e.

$$T^{-1} \sum_{t=1}^{T} f(w_t, z_t, \theta) \to E[f(w_t, z_t, \theta)] \text{ for } T \to \infty.$$

That requires IID or stationarity and weak dependence.

• If the moment conditions are correct, $g(\theta_0) = 0$, then GMM is consistent,

$$\hat{\theta}_{GMM}(W_T) \to \theta_0 \text{ as } T \to \infty,$$

for any $W_T$ positive definite.

• Intuition: If a LLN applies, then $g_T(\theta)$ converges to $g(\theta)$. Since $\hat{\theta}_{GMM}(W_T)$ minimizes the distance from $g_T(\theta)$ to zero, it will be a consistent estimator of the solution to $g(\theta_0) = 0$.

• The weight matrix, $W_T$, has to be positive definite, so that we put a positive and non-zero weight on all moment conditions.
Asymptotic Distribution

- Assume a central limit theorem for $f(w_t, z_t, \theta)$, i.e.:

$$\sqrt{T} \cdot g_T(\theta_0) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} f(w_t, z_t, \theta_0) \rightarrow N(0, S),$$

where $S$ is the asymptotic variance.

- Then it holds that for any positive definite weight matrix, $W$, the asymptotic distribution of the GMM estimator is given by

$$\sqrt{T} \left( \hat{\theta}_{GMM} - \theta_0 \right) \rightarrow N(0, V).$$

The asymptotic variance is given by

$$V = (D'WD)^{-1} D'WSWD (D'WD)^{-1},$$

where

$$D = E \left[ \frac{\partial f(w_t, z_t, \theta)}{\partial \theta'} \right]$$

is the expected value of the $R \times K$ matrix of first derivatives of the moments.
Efficient GMM Estimation

- The variance of $\hat{\theta}_{GMM}$ depends on the weight matrix, $W_T$.
  The efficient GMM estimator has the smallest possible (asymptotic) variance.

- Intuition: a moment with small variance is informative and should have large weight.
  It can be shown that the optimal weight matrix, $W_T^{opt}$, has the property that
  \[
  \text{plim } W_T^{opt} = S^{-1}.
  \]
  With the optimal weight matrix, $W = S^{-1}$, the asymptotic variance simplifies to
  \[
  V = (D'S^{-1}D)^{-1} D'S^{-1}SS^{-1}D (D'S^{-1}D)^{-1} = (D'S^{-1}D)^{-1}.
  \]

- The best moment conditions have small $S$ and large $D$.
  - A small $S$ means that the sample variation of the moment (noise) is small.
  - A large $D$ means that the moment condition is much violated if $\theta \neq \theta_0$.
    The moment is very informative on the true values, $\theta_0$.
    Related to the curvature of the criteria function as in ML.
• Hypothesis testing can be based on the asymptotic distribution:

\[ \hat{\theta}_{GMM} \sim N(\theta_0, T^{-1}\hat{V}) . \]

• An estimator of the asymptotic variance is given by

\[ \hat{V} = (D_T' S_T^{-1} D_T)^{-1} , \]

where

\[ \frac{D_T}{(R \times K)} = \frac{\partial g_T(\theta)}{\partial \theta'} = \frac{1}{T} \sum_{t=1}^{T} \frac{\partial f(w_t, z_t, \theta)}{\partial \theta'} \]

is the sample average of the first derivatives.
And \( S_T \) is an estimator of \( S = T \cdot V[g_T(\theta)] \). If the observations are independent, a consistent estimator is

\[ S_T = \frac{1}{T} \sum_{t=1}^{T} f(w_t, z_t, \theta) f(w_t, z_t, \theta)' . \]

Estimation of the weight matrix is typically the most tricky part of GMM.
Computational Issues

- The estimator is defined by minimizing $Q_T(\theta)$. Minimization can be done by

$$
\frac{\partial Q_T(\theta)}{\partial \theta} = \frac{\partial (g_T(\theta)' W_T g_T(\theta))}{\partial \theta} = 0 \quad (K \times 1).
$$

Sometimes analytically but often by numerical optimization.

- We need an optimal weight matrix, $W_T^{opt}$, but that depends on the parameters!

Two-step efficient GMM:

1. Choose an initial weight matrix, e.g. $W_{[1]} = I_R$, and find a consistent but inefficient first-step GMM estimator

$$
\widehat{\theta}_{[1]} = \arg \min_{\theta} g_T(\theta)' W_{[1]} g_T(\theta).
$$

2. Find the optimal weight matrix, $W_{[2]}^{opt}$, based on $\widehat{\theta}_{[1]}$. Find the efficient estimator

$$
\widehat{\theta}_{[2]} = \arg \min_{\theta} g_T(\theta)' W_{[2]}^{opt} g_T(\theta).
$$

The estimator is not unique as it depends on the initial weight matrix $W_{[1]}$. 
Iterated GMM estimator:

- From the estimator $\hat{\theta}_{[2]}$ it is natural to update the weights, $W^{opt}_{[3]}$, and update $\hat{\theta}_{[3]}$.

  We can switch between estimating $W^{opt}_{[\cdot]}$ and $\hat{\theta}_{[\cdot]}$ until convergence.

  Iterated GMM does not depend on the initial weight matrix.

  The two approaches are asymptotically equivalent.

Continuously updated GMM estimator:

- A third approach is to recognize from the outset that the weight matrix depends on the parameters, and minimize

  $$Q_T(\theta) = g_T(\theta)'W_T(\theta)g_T(\theta).$$

  That is never possible to solve analytically.
Test of Overidentifying Moment Conditions

• Recall that $K$ moment conditions are sufficient to estimate the $K$ parameters in $\theta$. If $R > K$, we can test the validity of the $R - K$ overidentifying moment conditions.

• By MM estimation we can set $K$ moment conditions equal to zero. If all $R$ conditions are valid then the $R - K$ moments should also be close to zero.

• From CLT we have

$$g_T(\theta_0) \overset{a}{\sim} N(0, T^{-1}S).$$

If we use the optimal weights, $W_T^{opt} \rightarrow S^{-1}$, then

$$\xi_J = T \cdot g_T(\hat{\theta}_{GMM})'W_T^{opt}g_T(\hat{\theta}_{GMM}) = T \cdot Q_T(\hat{\theta}_{GMM}) \rightarrow \chi^2(R - K).$$

• This is the J-test or the Hansen test for overidentifying restrictions. In linear models it is often referred to as the Sargan test. $\xi_J$ is not a test of the validity of model or the underlying economic theory. $\xi_J$ considers whether the $R - K$ moments are in line with the $K$ identifying moments.
Example: The C-CAPM Model

- Consider the consumption based capital asset pricing (C-CAPM) model of Hansen and Singleton (1982).

- A representative agent maximizes the discounted value of lifetime utility subject to a budget constraint:

\[
\max_{s=1}^{\infty} E \left[ \delta^s \cdot u(c_{t+s}) \mid \mathcal{I}_t \right],
\]

\[
A_{t+1} = (1 + r_{t+1}) A_t + y_{t+1} - c_{t+1},
\]

where \( A_t \) is financial wealth, \( y_t \) is income, \( 0 \leq \delta \leq 1 \) is a discount factor, and \( \mathcal{I}_t \) is the information set at time \( t \).

- The first order condition is given by the Euler equation:

\[
u'(c_t) = E \left[ \delta \cdot u'(c_{t+1}) \cdot R_{t+1} \mid \mathcal{I}_t \right],
\]

where \( u'(\cdot) \) is the derivative, and \( R_{t+1} = 1 + r_{t+1} \) is the return factor.
Now assume a constant relative risk aversion (CRRA) utility function:

\[ u(c_t) = \frac{c_t^{1-\gamma}}{1-\gamma}, \quad \gamma < 1, \]

so that \( u'(c_t) = c_t^{-\gamma} \). That gives the explicit Euler equation:

\[ c_t^{-\gamma} - E \left[ \delta \cdot c_{t+1}^{-\gamma} \cdot R_{t+1} \mid \mathcal{I}_t \right] = 0. \]

To ensure stationarity, we reformulate:

\[ E \left[ \delta \cdot \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \cdot R_{t+1} - 1 \mid \mathcal{I}_t \right] = 0, \]

which is a conditional moment condition.

That implies the unconditional moment conditions

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

for all variables \( z_t \in \mathcal{I}_t \) included in the formation set.
• To estimate the parameters, \( \theta = (\delta, \gamma)' \), we need at least \( R = 2 \) instruments in \( z_t \). We try with \( R = 3 \) instruments:

\[
z_t = \left( 1, \frac{c_t}{c_{t-1}}, R_t \right)'.
\]

• That produces the moment conditions

\[
E \left[ \left( \delta \cdot \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \cdot R_{t+1} - 1 \right) \right] = 0
\]

\[
E \left[ \left( \delta \cdot \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \cdot R_{t+1} - 1 \right) \left( \frac{c_t}{c_{t-1}} \right) \right] = 0
\]

\[
E \left[ \left( \delta \cdot \left( \frac{c_{t+1}}{c_t} \right)^{-\gamma} \cdot R_{t+1} - 1 \right) R_t \right] = 0,
\]

for \( t = 1, 2, \ldots, T \).

• The model is formally identified but \( \gamma \) is poorly determined. Weak instruments, little variation in the data, or wrong model!
Results for US data, $1959 : 3 - 1978 : 12$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Lags</th>
<th>$\delta$</th>
<th>$\gamma$</th>
<th>$T$</th>
<th>$\xi_J$</th>
<th>DF</th>
<th>$p - val$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-Step</td>
<td>$HC$</td>
<td>1</td>
<td>0.9987 (0.0086)</td>
<td>0.8770 (3.6792)</td>
<td>237</td>
<td>0.434</td>
<td>1</td>
</tr>
<tr>
<td>Iterated</td>
<td>$HC$</td>
<td>1</td>
<td>0.9982 (0.0044)</td>
<td>1.0249 (1.8614)</td>
<td>237</td>
<td>1.068</td>
<td>1</td>
</tr>
<tr>
<td>CU</td>
<td>$HC$</td>
<td>1</td>
<td>0.9981 (0.0044)</td>
<td>0.9549 (1.8629)</td>
<td>237</td>
<td>1.067</td>
<td>1</td>
</tr>
<tr>
<td>2-Step</td>
<td>$HAC$</td>
<td>1</td>
<td>0.9987 (0.0092)</td>
<td>0.8876 (4.0228)</td>
<td>237</td>
<td>0.429</td>
<td>1</td>
</tr>
<tr>
<td>Iterated</td>
<td>$HAC$</td>
<td>1</td>
<td>0.9980 (0.0045)</td>
<td>0.8472 (1.8757)</td>
<td>237</td>
<td>1.091</td>
<td>1</td>
</tr>
<tr>
<td>CU</td>
<td>$HAC$</td>
<td>1</td>
<td>0.9977 (0.0045)</td>
<td>0.7093 (1.8815)</td>
<td>237</td>
<td>1.086</td>
<td>1</td>
</tr>
<tr>
<td>2-Step</td>
<td>$HC$</td>
<td>2</td>
<td>0.9975 (0.0066)</td>
<td>0.0149 (2.6415)</td>
<td>236</td>
<td>1.597</td>
<td>3</td>
</tr>
<tr>
<td>Iterated</td>
<td>$HC$</td>
<td>2</td>
<td>0.9968 (0.0045)</td>
<td>−0.0210 (1.7925)</td>
<td>236</td>
<td>3.579</td>
<td>3</td>
</tr>
<tr>
<td>CU</td>
<td>$HC$</td>
<td>2</td>
<td>0.9958 (0.0046)</td>
<td>−0.5526 (1.8267)</td>
<td>236</td>
<td>3.501</td>
<td>3</td>
</tr>
<tr>
<td>2-Step</td>
<td>$HAC$</td>
<td>2</td>
<td>0.9970 (0.0068)</td>
<td>−0.1872 (2.7476)</td>
<td>236</td>
<td>1.672</td>
<td>3</td>
</tr>
<tr>
<td>Iterated</td>
<td>$HAC$</td>
<td>2</td>
<td>0.9965 (0.0047)</td>
<td>−0.2443 (1.8571)</td>
<td>236</td>
<td>3.685</td>
<td>3</td>
</tr>
<tr>
<td>CU</td>
<td>$HAC$</td>
<td>2</td>
<td>0.9952 (0.0048)</td>
<td>−0.9094 (1.9108)</td>
<td>236</td>
<td>3.591</td>
<td>3</td>
</tr>
</tbody>
</table>
Weight Matrix Estimation (Univariate Case)

- The optimal weight matrix is $S_T^{-1}$, where $S_T$ is a consistent estimator of

$$S = V[\sqrt{T} \cdot g_T(\theta)] = V \left[ \frac{\sqrt{T}}{T} \sum_{t=1}^{T} f_t \right] = \frac{1}{T} \cdot V \left[ \sum_{t=1}^{T} f_t \right].$$

- If $f_t$ and $f_s$ are independent, then the variance of the sum is the sum of the variances:

$$S = \frac{1}{T} \cdot V \left[ \sum_{t=1}^{T} f_t \right] = \frac{1}{T} \sum_{t=1}^{T} V[f_t] = \frac{1}{T} \sum_{t=1}^{T} E[f_t^2].$$

A natural estimator is

$$ST = \frac{1}{T} \sum_{t=1}^{T} f_t^2.$$

- This is robust to heteroskedasticity by construction and is often referred to as the heteroskedasticity consistent (HC) covariance estimator.
• If $f_t$ and $f_s$ are correlated, the variance includes the covariances:

$$S = \frac{1}{T} \cdot V \left[ \sum_{t=1}^{T} f_t \right] = \frac{1}{T} \left[ V(f_t) + 2 \cdot Cov(f_t, f_{t-1}) + 2 \cdot Cov(f_t, f_{t-2}) + \ldots \right].$$

• The heteroskedasticity and autocorrelation consistent (HAC) variance estimator is

$$S_T = \hat{V}(f_t) + \sum_{j=1}^{T-1} 2 \cdot \hat{Cov}(f_t, f_{t-j}),$$

where

$$\hat{Cov}(f_t, f_{t-j}) = \frac{1}{T} \sum_{t=j+1}^{T} f_t f_{t-j}.$$  

• Problems:

  1. We cannot estimate as many covariances as observations.
  2. The simple HAC estimator is not necessarily positive definite.
• We use a weight $w_j$ on covariance $j$, and let $w_j$ go to zero as $j$ increases. This class of so-called kernel estimators can be written as

$$S_T = \hat{V}(f_t) + \sum_{j=1}^{T-1} w_j \cdot 2 \cdot \hat{Cov}(f_t, f_{t-j}),$$

where $w_j = k\left(\frac{j}{B}\right)$. $k(\cdot)$ is a kernel function and $B$ is the bandwidth parameter.

• Example: Bartlett kernel (Newey-West estimator)
Example: 2SLS

- Consider again a regression model

\[ y_t = x'_t \beta_0 + \epsilon_t = x'_1 \gamma_0 + x'_2 \delta_0 + \epsilon_t, \]

where \( E[x_1 \epsilon_t] = 0 \) and \( E[x_2 \epsilon_t] \neq 0 \).

Assume that you have \( R > K \) valid instruments in \( z_t \) so that

\[ g(\beta_0) = E[z_t \epsilon_t] = E[z_t (y_t - x'_1 \beta_0)] = 0. \]

- The corresponding sample moments are given by

\[ g_T(\beta) = \frac{1}{T} \sum_{t=1}^{T} z_t (y_t - x'_t \beta) = \frac{1}{T} Z' (Y - X \beta), \]

where \( Y \ (T \times 1) \), \( X \ (T \times K) \), and \( Z \ (T \times R) \) are the stacked matrices.

- In this case we cannot solve \( g_T(\beta) = 0 \) directly; \( Z'X \) is \( R \times K \) and not invertible.
• Instead, we want to derive the GMM estimator by minimizing the criteria function

\[ Q_T(\beta) = g_T(\beta)' W_T g_T(\beta) \]

\[ = (T^{-1} Z' (Y - X\beta))' W_T (T^{-1} Z' (Y - X\beta)) \]

\[ = T^{-2} (Y' Z W_T Z' Y - 2\beta' X' Z W_T Z' Y + \beta' X' Z W_T Z' X \beta) . \]

• We take the first derivative, and the GMM estimator is the solution to

\[ \frac{\partial Q_T(\beta)}{\partial \beta} = -2T^{-2} X' Z W_T Z' Y + 2T^{-2} X' Z W_T Z' X \beta = 0. \]

We find \( \hat{\beta}_{GMM}(W_T) = (X' Z W_T Z' X)^{-1} X' Z W_T Z' Y \), depending on \( W_T \).

• To estimate the optimal weight matrix, \( W_T^{opt} = S_T^{-1} \), we use the estimator

\[ S_T = \frac{1}{T} \cdot \sum_{t=1}^{T} f(w_t, z_t, \theta) f(w_t, z_t, \theta)' = \frac{1}{T} \sum_{t=1}^{T} \tilde{\epsilon}_t^2 z_t z_t' , \]

which allows for general heteroskedasticity of the disturbance term.
• For the asymptotic distributions, we recall that

$$\hat{\beta}_{GMM} \sim N \left( \beta_0, T^{-1} \left( D' S^{-1} D \right)^{-1} \right).$$

The derivative is given by

$$D_T = \frac{\partial g_T(\beta)}{\partial \beta'} = \frac{\partial \left( T^{-1} \sum_{t=1}^{T} z_t (y_t - x_t' \beta) \right)}{\partial \beta'} = -T^{-1} \sum_{t=1}^{T} z_t x_t',$n

so the variance of the estimator becomes

$$V \left[ \hat{\beta}_{GMM} \right] = T^{-1} (D_T W_T^{opt} D_T)^{-1}$$

$$= T^{-1} \left( \left( -T^{-1} \sum_{t=1}^{T} x_t z_t' \right) \left( T^{-1} \sum_{t=1}^{T} \tilde{e}_t^2 z_t z_t' \right)^{-1} \left( -T^{-1} \sum_{t=1}^{T} z_t x_t' \right) \right)^{-1}$$

$$= \left( \sum_{t=1}^{T} x_t z_t' \right)^{-1} \sum_{t=1}^{T} \tilde{e}_t^2 z_t z_t' \left( \sum_{t=1}^{T} z_t x_t' \right)^{-1}. $$

• Note that this is the heteroskedasticity consistent (HC) variance estimator (White). GMM with allowance for heteroskedastic errors automatically produces heteroskedasticity consistent standard errors!
• If we assume that the error terms are IID, the optimal weight matrix simplifies to

\[ S_T = \hat{\sigma}^2 \frac{1}{T} \sum_{t=1}^{T} z_t z'_t = T^{-1} \hat{\sigma}^2 Z'Z, \]

where \( \hat{\sigma}^2 \) is a consistent estimator for \( \sigma^2 \).

• In this case the efficient GMM estimator becomes

\[
\hat{\beta}_{GMM} = (X'ZS_T^{-1}Z'X)^{-1} X'ZS_T^{-1}Z'Y.
\]

\[
= \left( X'Z \left( T^{-1} \hat{\sigma}^2 Z'Z \right)^{-1} Z'X \right)^{-1} X'Z \left( T^{-1} \hat{\sigma}^2 Z'Z \right)^{-1} Z'Y
\]

\[
= \left( X'Z (Z'Z)^{-1} Z'X \right)^{-1} X'Z (Z'Z)^{-1} Z'Y,
\]

which is identical to the two stage least squares (2SLS) estimator.

• The variance of the estimator is

\[ V \left[ \hat{\beta}_{GMM} \right] = T^{-1} (D'_T S_T^{-1} D_T)^{-1} = \hat{\sigma}^2 (X'Z (Z'Z)^{-1} Z'X)^{-1}, \]

which again coincides with the 2SLS variance.
Pseudo-ML (PML) Estimation

• The first order conditions for ML estimation can be seen as a sample counterpart to a moment condition

\[
\frac{1}{T} s(\theta) = \frac{1}{T} \sum_{t=1}^{T} s_t(\theta) = 0 \quad \text{corresponds to} \quad E[s_t(\theta)] = 0,
\]

and ML becomes a special case of GMM.

• \(\hat{\theta}_{ML}\) is consistent for weaker assumptions than maintained by ML. The FOC for a normal regression model corresponds to

\[
E[x_t(y_t - x_t'\beta)] = 0,
\]

which is weaker than the assumption that the entire distribution is correctly specified. OLS is consistent even if \(\epsilon_t\) is not normal.

• A ML estimation that maximizes a likelihood function different from the true models likelihood is referred to as a pseudo-ML or a quasi-ML estimator. Note that the variance matrix is no longer the inverse information.
# (My Unfair) Comparison of ML and GMM

<table>
<thead>
<tr>
<th>Assumptions:</th>
<th>Maximum Likelihood</th>
<th>Generalized Method of Moments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full specification.</td>
<td>Partial specification/weak assumptions.</td>
</tr>
<tr>
<td></td>
<td>Know Density($\theta_0$) apart from $\theta_0$.</td>
<td>Moment conditions: $E[f(\text{data};\theta_0)] = 0$.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strong economic assumptions.</td>
</tr>
<tr>
<td></td>
<td>(Smallest possible variance).</td>
<td>Larger than Cramér Rao.</td>
</tr>
<tr>
<td>Typical approach:</td>
<td>Statistical description of the data.</td>
<td>Estimate deep parameters of economic model.</td>
</tr>
<tr>
<td></td>
<td>Misspecification testing.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Restrictions recover economics.</td>
<td></td>
</tr>
<tr>
<td>Robustness:</td>
<td>First order conditions should hold!</td>
<td>Moment conditions should hold!</td>
</tr>
<tr>
<td></td>
<td>PML is a GMM interpretation of ML.</td>
<td>Weights and variances can be made robust.</td>
</tr>
<tr>
<td></td>
<td>Use larger PML variance.</td>
<td></td>
</tr>
</tbody>
</table>