GMM, Influence Functions, and Weight Matrices

AES Summer School in Structural Estimation

Toni M. Whited

July 2020

Why am I bothering to go over something this basic?

- When we do SMM we try to minimize something that looks like:

$$(\text{simulated moments} - \text{data moments})'(\text{weight matrix})(\text{simulated moments} - \text{data moments})$$

- Stephen just told you how to calculate the simulated moments
- The data moments are easy
- But how do you calculate the weight matrix?

Why am I bothering to go over something this basic?

- Far too many structural papers calculate weight matrices and standard errors incorrectly.
- Do not bootstrap the weight matrix. (Horowitz 2001).
- Calculating the weight matrix from simulated data is in principle fine, but you are taking the model too seriously.
- I am going to use basic GMM theory to teach you how to calculate weight matrices correctly and easily.
- Some of this material will be very new to you.

The Setup

- The following uses the notation in Wooldridge.
- Let
  - Let $\boldsymbol{w}_i$ be an $(M \times 1)$ be an $i.i.d.$ vector of random variables for observation i.
  - $\boldsymbol{\theta}$ be an $(P \times 1)$ vector of unknown coefficients.
  - $\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})$ be an $(L \times 1)$ vector of functions $\boldsymbol{g} : \left(\mathcal{R}^M \times \mathcal{R}^P\right) \rightarrow \mathcal{R}^L, \quad L \geq P$
- The function $\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})$ can be nonlinear.
- Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$.
- Let $\hat{\boldsymbol{\theta}}$ represent an estimate of $\boldsymbol{\theta}$.
- The “hat” and “naught” notation applies to anything we might want to estimate.

Moment Restrictions

- GMM is based on what are generally called moment restrictions and sometimes called orthogonality conditions (The latter terminology comes from the rational expectations literature.)

$$E\left(\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}_0\right)\right) = 0$$

- This condition is expressed in terms of the population. The corresponding sample moment restriction is

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) = 0$$

- What we want to do is choose $\hat{\boldsymbol{\theta}}$ to get $N^{-1}\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$ as close to zero as possible.

Examples of Moment Restrictions

- IV estimation:
  - Suppose you have a regression
  $$y_i = x_i\beta + u_i,$$
  and $E(u_i \mid x_i) \neq 0$.
  - **Or**, suppose you have a **nonlinear** regression
  $$y_i = f(x_i, \beta) + u_i,$$
  and $E(u_i \mid x_i) \neq 0$.
  - In **either** case suppose also that you have a vector of instruments z_i, that is uncorrelated with u_i, and whose dimension is at least as great as β. Then the moment restriction is
  $$E(z_i u_i) = 0$$

Criterion Function

- The estimator, $\hat{\boldsymbol{\theta}}$ minimizes a quadratic form:

$$\underset{}{\boldsymbol{Q}_N(\boldsymbol{\theta})} = \underset{(1\times L)}{\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}(\boldsymbol{w}_i,\boldsymbol{\theta})\right]'}\underset{(L\times L)}{\widehat{\boldsymbol{\Xi}}}\underset{(L\times 1)}{\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}(\boldsymbol{w}_i,\boldsymbol{\theta})\right]}$$

where $\widehat{\boldsymbol{\Xi}}$ is a positive definite matrix that converges in probability to $\boldsymbol{\Xi}_0$

- In this case, $\boldsymbol{Q}_N$ converges in probability to

$$\{E[\boldsymbol{g}(\boldsymbol{w}_i,\boldsymbol{\theta})]\}'\boldsymbol{\Xi}\{E[\boldsymbol{g}(\boldsymbol{w}_i,\boldsymbol{\theta})]\}$$

Exact and Overidentification

- If $L = P$, then the estimator is exactly identified, and we can find $\boldsymbol{\theta}$ by solving

$$N^{-1} \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) = \mathbf{0}$$

- If $L > P$, the model is overidentified and if it is nonlinear, you usually have to use numerical techniques.

- If $\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$ has first derivatives with no closed form solutions, these numerical techniques can take a very long time.

Optimal Weighting Matrix

- The symbol $\boldsymbol{\Xi}$ represents any arbitrary, positive definite weighting matrix.
- The optimal weighting matrix is the inverse of the variance of $\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})$. Call this variance

$$\boldsymbol{\Lambda} \equiv E\left(\boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})\, \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})'\right).$$

- Estimating $\widehat{\boldsymbol{\Lambda}}$. Doing GMM is a bit circular. We want to minimize

$$\boldsymbol{Q}_N(\boldsymbol{\theta}) = \left[N^{-1}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})\right]' \widehat{\boldsymbol{\Lambda}}^{-1} \left[N^{-1}\sum_{i=1}^{N} \boldsymbol{g}(\boldsymbol{w}_i, \boldsymbol{\theta})\right]$$

to get an estimate of $\boldsymbol{\theta}$. But we need an estimate of $\boldsymbol{\theta}$ to estimate $\widehat{\boldsymbol{\Lambda}}$.

Estimating the Optimal Weighting Matrix

- You can estimate $\hat{\Lambda}$ by

$$\widehat{\boldsymbol{\Lambda}} \equiv \frac{1}{N} \sum_{i=1}^{N} \left[\boldsymbol{g} \left(\boldsymbol{w}_i, \boldsymbol{\theta} \right) \right] \left[\boldsymbol{g} \left(\boldsymbol{w}_i, \boldsymbol{\theta} \right) \right]'$$

- The usual procedure is as follows:
  - Estimate $\boldsymbol{\theta}$ using $\widehat{\boldsymbol{\Lambda}} \equiv I$. (This $\boldsymbol{\theta}$ is consistent but not efficient.)
  - Use this estimate of $\boldsymbol{\theta}$ to estimate $\widehat{\boldsymbol{\Lambda}}$.
  - Re-estimate $\boldsymbol{\theta}$ using the estimate of $\widehat{\boldsymbol{\Lambda}}$.
  - Keep going until $\boldsymbol{\theta}$ converges.
  - With most SMM applications, you can calculate the weight matrix without knowledge of the model parameters, so all of this is unnecessary.

Asymptotic Distribution of GMM Estimators

- Define the following

$$\begin{aligned} \boldsymbol{G}' &= \left.\frac{\partial \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_N} \\ \boldsymbol{G}_0' &= E\left(\boldsymbol{G}'\right) \end{aligned}$$

  - Note $\partial \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) / \partial \boldsymbol{\theta}'$ is a MATRIX

$$\partial \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) / \partial \boldsymbol{\theta}' \equiv \left[\begin{array}{llll} \partial g_1 / \partial \theta_1 & \partial g_1 / \partial \theta_2 & \ldots & \partial g_1 / \partial \theta_P \\ \vdots & \vdots & \ddots & \vdots \\ \partial g_L / \partial \theta_1 & \partial g_L / \partial \theta_2 & \ldots & \partial g_L / \partial \theta_P \end{array}\right]$$

  - The Jacobian of $\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$ w.r.t. $\boldsymbol{\theta}$, with dimension $P \times L$.

Asymptotic Distribution of GMM Estimators

- Then the asymptotic distribution of $\sqrt{N}\left(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)$ is $N\left(0,\boldsymbol{V}\right)$, in which

$$\begin{aligned}\boldsymbol{V} \quad &\equiv \quad \left[\boldsymbol{G}_0\boldsymbol{\Lambda}^{-1}\boldsymbol{G}_0'\right]^{-1} \\ & \qquad (P\times L)(L\times L)(L\times P)\end{aligned}$$

- Heuristic "Proof:" Recall that we are trying to minimize:

$$\boldsymbol{Q}_N\left(\boldsymbol{\theta}\right)=\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right]'\widehat{\boldsymbol{\Xi}}\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right]$$

- How do you minimize anything? Take the derivative w.r.t $\boldsymbol{\theta}$ and set the result equal to zero.

Asymptotic Distribution of GMM Estimators

- Take the derivative, and w.p.a. 1, it equals zero.

$$
\begin{aligned}
\partial \boldsymbol{Q}_N\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) / \partial \boldsymbol{\theta} &= 0 \\
2\left[N^{-1} \sum_{i=1}^{N} \boldsymbol{G}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right] \widehat{\boldsymbol{\Xi}}\left[N^{-1} \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right] &= 0 \\
\left[N^{-1} \sum_{i=1}^{N} \boldsymbol{G}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right] \widehat{\boldsymbol{\Xi}}\left[N^{-1} \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)\right] &= 0
\end{aligned}
$$

- Now take a **mean**-value (not Taylor) expansion of $\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$

$$
\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) = \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) + \sum_{i=1}^{N} \boldsymbol{G}'\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right)
$$

in which $\bar{\theta}$ is some vector between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$.

Asymptotic Distribution of GMM Estimators

- Now we substitute this mean value expansion into the first order condition, replace random averages with their plims, and solve away.

$$
\begin{aligned}
\boldsymbol{G}_0 \boldsymbol{\Xi}_0 \left[N^{-1} \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) + \boldsymbol{G}_0' \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) \right] &= 0 \\
\boldsymbol{G}_0 \boldsymbol{\Xi}_0 \boldsymbol{G}_0' \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) &= -\boldsymbol{G}_0 \boldsymbol{\Xi}_0 \left[\sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) \right] \\
\left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) &= -\left(\boldsymbol{G}_0 \boldsymbol{\Xi}_0 \boldsymbol{G}_0'\right)^{-1} \boldsymbol{G}_0 \boldsymbol{\Xi}_0 \left[N^{-1} \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) \right] \\
\sqrt{N} \left(\boldsymbol{\theta} - \boldsymbol{\theta}_0\right) &= -\left(\boldsymbol{G}_0 \boldsymbol{\Xi}_0 \boldsymbol{G}_0'\right)^{-1} \boldsymbol{G}_0 \boldsymbol{\Xi}_0 \left[N^{-1/2} \sum_{i=1}^{N} \boldsymbol{g}\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) \right]
\end{aligned}
$$

- The expression in this last line contains what is called an “influence function.” We will come back to this shortly.

Asymptotic Distribution of GMM Estimators

- So, dropping the "naught" notation from G and Ξ, for simplicity:

$$\begin{aligned} E\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)' &\equiv \\ E\left\{\left(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}'\right)^{-1}\boldsymbol{G}\boldsymbol{\Xi}\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right)\right]\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right)\right]'\boldsymbol{\Xi}\boldsymbol{G}'\left(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}'\right)^{-1}\right\} &= \\ \left(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}'\right)^{-1}\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{\Lambda}\boldsymbol{\Xi}\boldsymbol{G}'\left(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}'\right)^{-1} & \end{aligned}$$

- Note that if we set $\boldsymbol{\Xi}\equiv\boldsymbol{\Lambda}^{-1}$, then this mess reduces as follows:

$$\begin{aligned} \left(\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\right)^{-1}\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\left(\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\right)^{-1} &= \\ \left(\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\right)^{-1}\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\left(\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\right)^{-1} &= \\ \left(\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}'\right)^{-1} & \end{aligned}$$

Asymptotic Distribution of GMM Estimators

- $(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}')^{-1}\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{\Lambda}\boldsymbol{\Xi}\boldsymbol{G}'(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}')^{-1}$ is always greater than $(\boldsymbol{G}\boldsymbol{\Lambda}^{-1}\boldsymbol{G}')^{-1}$, in the sense that the difference between the two is a positive definite matrix.
- So an efficient estimate of the variance of $\widehat{\boldsymbol{\theta}}$ is given by

$$\frac{1}{N}\left\{\boldsymbol{G}\widehat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{G}'\right\}^{-1}$$

- What if we want to calculate the variance of an $H \times 1$ dimensional function $\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}\right)$, $H \leq P$?
- We can use the “delta-method,” which gives the variance of $\boldsymbol{h}\left(\widehat{\boldsymbol{\theta}}\right)$ as

$$\left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}\right)\left\{\frac{1}{N}\left\{\boldsymbol{G}\widehat{\boldsymbol{\Lambda}}^{-1}\boldsymbol{G}'\right\}^{-1}\right\}\left(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}\right)'$$

Overidentifying Restrictions

- If $L > P$, the model is overidentified. We have more equations than unknowns.
- Presumably we could take different subsets of P equations and solve exactly for the P elements of $\boldsymbol{\theta}$.
- Testing the overidentifying restrictions intuitively is a matter of testing to see if different exactly identified subsets of moment restrictions have the same solution.
- If the model is correct, then each of these answers should be the same.
  - What does this idea tell you about an informal way to see if the overidentifying restrictions are rejected?

Hansen's J-Test

- The following statistic

$$N\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right)'\widehat{\boldsymbol{\Lambda}}^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right)$$

  converges to a χ^2 statistic with $(L-P)$ degrees of freedom under the null that the overidentifying restrictions hold.

- This is what is called a portmanteau test. "Wearovercoat"

- It tests for general misspecification, not any specific sort.

- The GMM J-test therefore need not be very powerful to detect misspecification.

Why am I torturing you with influence functions?

- Our GMM/SMM weight matrices will be very complicated. Example:
  - Moments = M = [Mean Variance RegressionSlope1 RegressionSlope2]
  - Optimal weight matrix = cov$(M)^{-1}$
  - How to estimate covariance between a mean and a variance?
  - How to estimate covariance between slopes from two separate regressions?
  - Etc.
  - Especially hard if we need to cluster by firm, etc.
- Influence functions give you a simple way to estimate cov(M)
- Beyond structural estimation, if you know how to use influence functions, **you know how to estimate the standard error for anything!**
  - Test for serial correlation of the residuals of a nonlinear panel model?
  - Hausman test when the Hausman assumptions are not satisfied.

General Definition of an Influence Function

- Consider any estimator $\hat{\theta}$, and suppose there is a function $\phi(\boldsymbol{w}_i)$ such that

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=\sum_{i=1}^{N}\phi(\boldsymbol{w}_i)/\sqrt{N}+o_p(1), \quad E(\phi(\boldsymbol{w}_i)=0, \quad E(\phi(\boldsymbol{w}_i)\phi(\boldsymbol{w}_i)') \text{ exists.}$$

- then $\phi(\boldsymbol{w}_i)$ is called the influence function of $\hat{\theta}$.
- In words, it gives the effect of a single observation on the estimator, up to the $o_p(1)$ remainder term.
- In different words, an influence function is a function of the data whose mean has the same asymptotic variance as the estimator.

Influence Function for a GMM Estimator

- Now reconsider the expression

$$\sqrt{N}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)=-\left(\boldsymbol{G}_0\boldsymbol{\Xi}_0\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\boldsymbol{\Xi}_0\left[N^{-1/2}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right)\right]$$

  and compare it to the general expression for an influence function:

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right)=\sum_{i=1}^{n}\phi(\boldsymbol{w}_i)/\sqrt{N}+o_p(1)$$

- The only real difference is the $o_p(1)$ term, but somewhere in there we substituted plims in for sample averages, so the influence function for a GMM estimator must be:

- So the influence function for a GMM estimator must be:

$$-\left(\boldsymbol{G}_0\boldsymbol{\Xi}_0\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\boldsymbol{\Xi}_0\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right)$$

- End of proof by staring. For a real, but logically similar proof, see Newey and McFadden (1994).

Example

- Recall the definition of an influence function (and dropping the zero subscripts):

$$-\left(\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{G}'\right)^{-1}\boldsymbol{G}\boldsymbol{\Xi}\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right)$$

- What is the influence function for the estimate of the mean of a random variable z_i with mean μ?

$$\begin{aligned}
\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right) &\equiv z_i-\mu \\
\boldsymbol{\Xi} &\equiv 1 \\
\boldsymbol{G} &\equiv 1 \\
\phi\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right) &\equiv -\left(z_i-\mu\right)
\end{aligned}$$

The sample counterpart for observation i is

$$-z_i+N^{-1}\sum_{i=1}^{N}z_i$$

Example

- Consider a simple linear regression

$$y_i = x_i\beta + u_i$$

- What is the influence function for β?

$$\begin{aligned}
\boldsymbol{g}\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) &\equiv x_i \cdot u_i \\
&= x_i \cdot (y_i - x_i\beta) \\
\boldsymbol{\Xi} &\equiv \sigma^2 E\left(x_i' x_i\right)^{-1} \\
\boldsymbol{G} &\equiv E\left(x_i' x_i\right) \\
\phi\left(\boldsymbol{w}_i, \bar{\boldsymbol{\theta}}\right) &\equiv -E\left(x_i' x_i\right)^{-1}\left(x_i \cdot (y_i - x_i\beta)\right)
\end{aligned}$$

The sample counterpart for observation i is

$$-\left(N^{-1}\sum_{i=1}^{N}\left(x_i' x_i\right)\right)^{-1}\left(x_i \cdot u_i\right)$$

where the operator $\cdot$ is the Hadamard element-by-element operator.

Stacking

- What if you estimate the mean μ and the OLS coefficient β, and you want to know the covariance between these two estimates?
- Option 1: Just estimate them jointly in a big GMM system.
- Option 2: Stack the influence functions and take the inner product.[1]
- Let $\hat{\phi}_{\mu}$ be the $N \times 1$ sample influence function for μ.
- Let $\hat{\phi}_{\beta}$ be the $N \times k$ sample influence function for β.

[1] The reference for this is Erickson and Whited (2002).

Stacking

- Let's define

$$\Phi_{\mu\beta} \equiv \left[\left(-z + N^{-1}\sum_{i=1}^{N} z_i\right) \qquad \left(-\left(N^{-1}\sum_{i=1}^{N}(x_i' x_i)\right)^{-1}(x \cdot u)\right) \right]$$

- Notice I dropped the i subscripts.
- The dimension of this matrix is $N \times (k+1)$.
- The sample covariance matrix for $\begin{pmatrix} \mu \\ \beta \end{pmatrix}$ is then

$$\Phi'_{\mu\beta}\Phi_{\mu\beta}N^{-2}$$

Sample Matlab Code

```
% Mean influence function
n = size(z);
meaninflnc = z - mean(z).*ones(n,1);

% OLS influence function
b=inv(x'*x)*x'*y;
uhat = y - x*b;
olsinflnc=inv((x'*x)./n)*((x.*uhat(:,ones(size(x,2),1))));

% Big influence function
biginflnc = zeros(n,size(x,2)+1);
biginflnc(:,1) = meaninflnc;
biginflnc(:,2:(size(x,2)+1)) = olsinflnc;

% Covary the influence functions
avar = biginflnc'*biginflnc./(n^2);
```

Two-Step Estimation

- Suppose you are doing a GMM estimator, but you estimate one or more of the parameters separately via a different procedure, and then plug these estimates into your GMM moment equations.
- Why? Sometimes this type of exercise reduces the dimensionality of the problem substantially.
- How do you figure out the GMM covariance matrix?
- This is nontrivial because the GMM estimates inherit the sampling variability from the first step.

Two-Step Estimation

- Let δ be a parameter vector of dimension S that you estimate in a first step via a different procedure
- Then you plug δ into your moment vector to get

$$\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_i, \boldsymbol{\delta})$$

  and use this moment vector to estimate $\boldsymbol{\theta}$.
- The variance of the two-step estimator is

$$\left(\boldsymbol{G}\boldsymbol{\Omega}^{-1}\boldsymbol{G}'\right)^{-1}$$

- You can estimate Ω by

$$\widehat{\boldsymbol{\Omega}} \equiv \frac{1}{N}\sum_{i=1}^{N}\left[\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) - \mathbb{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_i, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)\phi^{\boldsymbol{\delta}}\left(\boldsymbol{\delta}, \boldsymbol{w}_i\right)\right]\left[\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) - \mathbb{E}\left(\frac{\partial \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_i, \boldsymbol{\delta})}{\partial \boldsymbol{\delta}}\right)\phi^{\boldsymbol{\delta}}\left(\boldsymbol{\delta}, \boldsymbol{w}_i\right)\right]'$$

  in which $\phi^{\boldsymbol{\delta}}$ is the influence function for $\boldsymbol{\delta}$.
- A clear derivation of this estimator is in Newey and McFadden's chapter in the 4^{th} volume of the *Handbook of Econometrics*.

Clustered Weight Matrices

- Everything I have taught you thus far is for $i.i.d.$ data. Data are almost never $i.i.d.$ in corporate finance (or accounting).
- So how do you calculate a weight matrix and get your standard errors right if the data are not $i.i.d$?
- We will consider the following case.
  - The sample consists of K groups (clusters) of n_k observations each $(N = n_1 + \cdots + n_K)$
  - Observations are independent across groups but dependent within groups
  - $K \to \infty$, and n_k fixed for each k.

Clustered Weight Matrices

- We order observations by groups and use double-index notation so that

$$\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}) \equiv \{\boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_{1,1}), \ldots, \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_{n_1,1}) \mid \ldots \mid \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_{1,K}), \ldots, \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_{n_k,K})\}$$

- Under cluster sampling, the observations $\boldsymbol{w}_{n,k}$ might be dependent within a cluster, k.

- I'm going to simplify notation

$$\begin{aligned} \boldsymbol{g}_{1,1} &\equiv \boldsymbol{g}(\boldsymbol{\theta}, \boldsymbol{w}_{1,1}) \\ \hat{\boldsymbol{g}}_{1,1} &\equiv \boldsymbol{g}(\hat{\boldsymbol{\theta}}, \boldsymbol{w}_{1,1}) \end{aligned}$$

Clustered Weight Matrices

- Let

$$\bar{\boldsymbol{g}} = \sum_{j=1}^{n_k} \boldsymbol{g}_{j,k}$$

- Then we can define Λ as:

$$\Lambda = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{K} E\left(\bar{\boldsymbol{g}}_k \bar{\boldsymbol{g}}_k'\right).$$

- Note that $E\left(\bar{\boldsymbol{g}}_i \bar{\boldsymbol{g}}_j'\right) = 0$ only if i and j belong to different clusters.
- Define:

$$\tilde{\boldsymbol{g}} = \sum_{j=1}^{n_k} \hat{\boldsymbol{g}}_{j,k}$$

- A consistent estimate of Λ is therefore:

$$\hat{\Lambda} = \frac{1}{N} \sum_{k=1}^{K} \tilde{\boldsymbol{g}}_k \tilde{\boldsymbol{g}}_k'.$$

Sample Matlab code for a balanced panel

Consider a panel with dimensions T and N:

```
k = size(biginflnc,2);
vmx = zeros(k,k);
for qq=1:capN;
    phii = sum(biginflnc((qq-1)*(capT)+1:qq*(capT),:));
    vmx = vmx + phii'*phii;
end;

vmx = vmx./((capN*capT)^2);
```

Dynamic models require the estimation of dynamics

- A large fraction of dynamic models have driving processes that follow autoregressive processes.
- One statistic that needs to be matched in many structural estimations is an $AR(1)$ coefficient.

Consistent estimation of a first-order autoregressive coefficient with fixed effects

- Suppose you have a variable

$$y_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T.$$

  that follows a process

$$y_{it} = \alpha_i + \rho y_{it-1} + u_{it},$$

  in which u_{it} is possibly correlated with α_i.

- OLS will not work.
- You cannot do firm-level deviations from means with a lagged dependent variable.
- Dynamic panel models suffer from weak instrument problems.

Han and Phillips (2010)

- Han and Phillips (2010) use double differencing to remove the fixed effect.
- Some ugly but easy algebra shows that a consistent estimate of ρ is given by

$$\hat{\rho} = \frac{\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta y_{it-1} \left(2\Delta y_{it} + \Delta y_{it-1}\right)}{\sum_{i=1}^{N} \sum_{t=2}^{T} \left(\Delta y_{it-1}\right)^2}$$

  where Δ is the first difference operator.
- This estimator is clearly obtained from regressing $2\Delta y_{it} + \Delta y_{it-1}$ on Δy_{it-1}.

Albuquerque, R.„ Schroth, E., 2010. Quantifying private benefits of control from a structural model of block trades. Journal of Financial Economics 96, 33–55.

Claessens, S., Djanov, S., Fan, J.P.H.„ Lang, L.H.P., 2002. Disentangling the incentive and entrenchment effects of large shareholdings. Journal of Finance 57, 2741–2771.

Erickson, T.„ Whited, T.M., 2002. Two-step GMM estimation of the errors-in-variables model using high-order moments. Econometric Theory 18, 776–799.

Fazzari, S.M., Hubbard, R.G.„ Petersen, B.C., 1988. Financing constraints and corporate investment. Brookings Papers on Economic Activity 1, 141–206.

Han, C.„ Phillips, P.C.B., 2010. GMM estimation for dynamic panels with fixed effects and strong instruments at unity. Econometric Theory 26, 119–151.

Hansen, L.P.„ Singleton, K.J., 1982. Generalized instrumental variables estimation of nonlinear rational expectations models. Econometrica 50, 1269–1286.

Horowitz, J.L., 2001. The bootstrap, in: Heckman, J.J.„ Leamer, E. (Eds.), Handbook of Econometrics. Elsevier. volume 5 of *Handbook of Econometrics*, pp. 3159 – 3228.

Newey, W.„ McFadden, D., 1994. Large sample estimation and hypothesis testing, in: Engle, R.„ McFadden, D. (Eds.), Handbook of Econometrics, Vol. 4. North-Holland, Amsterdam, pp. 2111–2245.

Whited, T.M., 1992. Debt, liquidity constraints, and corporate investment: Evidence from panel data. Journal of Finance 47, 1425–1460.