Summer School in Structural Estimation
GMM, Influence Functions, and Weight Matrices

Toni M. Whited

June 2023

Outline

Why am I bothering to go over something as basic as GMM?

- We are going to be solving simulated moments problems that look like this:

- The simulated moments estimator of $\boldsymbol{\theta}$ is defined as the solution to the minimization of

$$\begin{aligned}
\hat{\boldsymbol{\theta}} &= \arg\min_{\theta} \boldsymbol{Q}_n(\boldsymbol{\theta}) \equiv \left[n^{-1}\sum_{i=1}^{\infty}\left(h(\boldsymbol{w}_i) - S^{-1}\sum_{s=1}^{S} h(y_{is}(\boldsymbol{\theta}))\right)\right]' \hat{\boldsymbol{\Xi}}_n \left[n^{-1}\sum_{i=1}^{\infty}\left(h(\boldsymbol{w}_i) - S^{-1}\sum_{s=1}^{S} h(y_{is}(\boldsymbol{\theta}))\right)\right] \\
&= \arg\min_{\theta} \boldsymbol{Q}_n(\boldsymbol{\theta}) \equiv [\text{data moments} - \text{simulated moments}(\boldsymbol{\theta})]' \hat{\boldsymbol{\Xi}}_n [\text{data moments} - \text{simulated moments}(\boldsymbol{\theta})]
\end{aligned}$$

- $\hat{\boldsymbol{\Xi}}_n$ is a positive definite matrix that converges in probability to a deterministic positive definite matrix $\boldsymbol{\Xi}$.

We want to find the most efficient version of an estimate of Ξ

- It turns out that this estimator is also an example of a GMM estimator.
- But why do we care about the estimate of the weight matrix?

It is important to calculate the weight matrix correctly

- Far too many structural papers calculate weight matrices and standard errors incorrectly.
- Do not bootstrap the weight matrix. (Horowitz 2001). It is consistent but can be very biased in finite samples.
- Calculating the weight matrix from simulated data doesn't take into account sampling variation in the actual data.
- I am going to use basic GMM theory to teach you how to calculate weight matrices correctly and easily.
- We will be stacking influence functions, and this will be new to most of you.

Outline

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.
  - $\boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right)$ 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, $\hat{\theta}$ 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 usually 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 $\hat{\mathbf{\Lambda}}$ by

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

- The usual procedure is as follows:
  - Estimate $\boldsymbol{\theta}$ using $\widehat{\mathbf{\Lambda}} \equiv I$. (This $\boldsymbol{\theta}$ is consistent but not efficient.)
  - Use this estimate of $\boldsymbol{\theta}$ to estimate $\widehat{\mathbf{\Lambda}}$.
  - Re-estimate $\boldsymbol{\theta}$ using the estimate of $\widehat{\mathbf{\Lambda}}$.
  - Keep going until $\boldsymbol{\theta}$ converges.

Estimating the Optimal Weighting Matrix

- Researches used to use a two-step procedure.
- However, after many Monte Carlo studies over the years, most now iterate.
- Another possibility is just to minimize Q_N all at once.
- If the application permits, sometimes the optimal weighting matrix does not depend on unknown parameters, and no iteration is necessary.
- This is the case with the SMM estimators we are using in this class.

GMM Influence Function Outline

- We will derive (informally) the asymptotic distribution of a GMM estimator.
- We will do this by linearizing so that the GMM estimator is not the argmax of a complicated function.
- Instead, we will express it as a closed-form function up to a term that converges in probability to zero.
- That term is the influence function

Asymptotic Distribution of GMM Estimators

- Define the following

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

  - Note $\boldsymbol{G}' \equiv \partial \boldsymbol{g}\left(\boldsymbol{w}_i, \boldsymbol{\theta}\right) / \partial \boldsymbol{\theta}'$ is the Jacobian matrix and is a function of the data

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

  - G had dimension $P \times L$, and G' had dimension $L \times P$.

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

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

- Derivation: 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 set it to zero.

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

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

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

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.
- First order condition

$$\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{G}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right]\widehat{\boldsymbol{\Xi}}\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right]=0$$

- Mean value expansion

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

- Substitution

$$\left[N^{-1}\sum_{i=1}^{N}\boldsymbol{G}\left(\boldsymbol{w}_i,\boldsymbol{\theta}\right)\right]\widehat{\boldsymbol{\Xi}}\left[N^{-1}\left(\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right)+\sum_{i=1}^{N}\boldsymbol{G}'\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)\right)\right]=0$$

Asymptotic Distribution of GMM Estimators

- Now replace random averages with their plims.
- So there should be an $o_p(1)$ term floating around.

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

and solve away.

Asymptotic Distribution of GMM Estimators

- Solving away . . .

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

- The right-hand side of the second-to-last line contains what is called an “influence function.”

Asymptotic Distribution of GMM Estimators

- So the variance of the GMM estimator can be obtained by covarying the influence functions!

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

- Note that if we set $\boldsymbol{\Xi}_0 \equiv \boldsymbol{\Lambda}_0^{-1}$, then this mess reduces as follows:

$$
\begin{aligned}
\left(\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{\Lambda}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\left(\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\right)^{-1} &= \\
\left(\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\left(\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\right)^{-1} &= \\
\left(\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\right)^{-1}
\end{aligned}
$$

Asymptotic Distribution of GMM Estimators

- $\left(\boldsymbol{G}_0\boldsymbol{\Xi}_0\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\boldsymbol{\Xi}_0\boldsymbol{\Lambda}_0\boldsymbol{\Xi}_0\boldsymbol{G}_0'\left(\boldsymbol{G}_0\boldsymbol{\Xi}_0\boldsymbol{G}_0'\right)^{-1}$ is always greater than $\left(\boldsymbol{G}_0\boldsymbol{\Lambda}_0^{-1}\boldsymbol{G}_0'\right)^{-1}$,
- 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

$$\frac{1}{N}\left\{\widehat{\boldsymbol{G}_0}\widehat{\boldsymbol{\Lambda}}^{-1}\widehat{\boldsymbol{G}_0}'\right\}^{-1}$$

Delta Method

- What if we want to calculate the variance of an $R \times 1$ dimensional function $\boldsymbol{r}\left(\widehat{\boldsymbol{\theta}}\right)$, $R \leq P$?
- We can use the "delta method," which gives the variance of $\boldsymbol{r}\left(\widehat{\boldsymbol{\theta}}\right)$.
- Informal derivation using a Taylor expansion:

$$\boldsymbol{r}\left(\widehat{\boldsymbol{\theta}}\right) \approx \boldsymbol{r}\left(\boldsymbol{\theta}_0\right) + \left(\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\theta}}\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)$$

- So

$$\text{var}\left(r\left(\widehat{\boldsymbol{\theta}}\right)\right) \approx \left(\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\theta}}\right)\left\{\frac{1}{N}\left\{\boldsymbol{G}_0 \widehat{\boldsymbol{\Lambda}_0}^{-1} \boldsymbol{G}_0'\right\}^{-1}\right\}\left(\frac{\partial \boldsymbol{r}}{\partial \boldsymbol{\theta}}\right)'$$

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.

- I will talk about the power of the test in the context of SMM later.

Outline

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.
- An estimator that has an influence function is called an asymptotically linear estimator.

Slightly More Formal Stuff

- Consider an estimator of a real parameter $\theta \in \Theta$, where Θ is an open convex subset of $\mathbb{R}$, based on a sample w_N of size N.
- Consider a family $\mathcal{F}$ of distributions $\{F_\theta : \theta \in \Theta\}$.
- Consider estimators $\hat{\theta} = T(\hat{F}_N)$, where $\hat{F}_N$ is the empirical distribution and $T(\cdot)$ is a functional.
  - Example, a mean, μ, is given by $\int w dF$, which is a functional.
- All necessary differentiability and boundedness assumptions are satisfied.

Slightly More Formal Stuff

- Let δ_w be a distribution with a mass of 1 $(\leq w)$, and zero elsewhere.
- The influence function of T at F is a special case of a Gâteaux derivative:

$$\phi(w; T, F) \equiv \lim_{\epsilon \to 0} \frac{T((1-\epsilon)F + \epsilon\delta_w) - T(F)}{\epsilon}$$

- It puts a smidge of extra weight on one observation. Hence, the name "influence function."
- A Gâteaux derivative is intuitively the analogue of a directional derivative from multivariable calculus to function spaces.

Heuristics as N gets large

- Let's call $G \equiv T((1-\epsilon)F + \epsilon\delta_w)$. Then we very roughly have a "Taylor series" type of result:

$$T(G) = T(F) + \int \phi(w; T, F) d(G - F)(w) + \text{remainder}$$

- As N gets large, the empirical distribution $\hat{F}_N$ tends to the theoretical distribution, F (Glivenko-Cantelli), and $T(\hat{F}_N)$ tends to $T(F)$

- Note that the empirical distribution is just $\hat{F}_N = N^{-1} \sum_{i=1}^{N} \delta_{w_i}$

- So

$$\begin{aligned} T(\hat{F}_N) - T(F) &\approx N^{-1} \sum_{i=1}^{N} \phi(w_i; T, F) + \text{remainder} \\ \sqrt{N}(T(\hat{F}_N)) - T(F)) &\approx N^{-\frac{1}{2}} \sum_{i=1}^{N} \phi(w_i; T, F) + \text{remainder} \end{aligned}$$

Influence Function for a GMM Estimator

- Now reconsider the expression

$$\sqrt{N}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)=-\left(\boldsymbol{G}_0\Xi_0\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\Xi_0\left[N^{-1/2}\sum_{i=1}^{N}\boldsymbol{g}\left(\boldsymbol{w}_i,\hat{\boldsymbol{\theta}}\right)\right]+o_p(1)$$

and compare it to the general expression for an influence function:[1]

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

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

$$-\left(\boldsymbol{G}_0\Xi_0\boldsymbol{G}_0'\right)^{-1}\boldsymbol{G}_0\Xi_0\boldsymbol{g}\left(\boldsymbol{w}_i,\hat{\boldsymbol{\theta}}\right)$$

- End of proof by staring. For a real, extremely brief, but logically similar proof, see Newey and McFadden (1994). You can also derive it from the formal definition.

[1] I replaced the bar on $\boldsymbol{\theta}$ with a hat because the difference ends up in the $o_p(1)$ term.

Example

- Recall the definition of an influence function :

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

- What is the influence function for the estimate of the mean of a random variable z_i with mean μ and variance σ^2?

$$\begin{aligned}
\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right) &\equiv z_i-\mu \\
\boldsymbol{\Xi} &\equiv \sigma^{-2} \\
\boldsymbol{G} &\equiv -1 \\
\phi\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right) &\equiv \left(z_i-\mu\right)
\end{aligned}$$

The sample counterpart for observation i is

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

Example

- Recall the definition of an influence function :

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

- Consider a simple linear regression

$$y_i = x_i\beta + u_i$$

- What is the influence function for β?

$$\begin{aligned}
\boldsymbol{g}\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right) &\equiv x_i \cdot u_i = x_i \cdot (y_i - x_i\beta) \\
\boldsymbol{\Xi}_0 = \boldsymbol{\Lambda}_0^{-1} &\equiv \sigma^{-2}E\left(x_i'x_i\right)^{-1} \\
\boldsymbol{G}_0 &\equiv -E\left(x_i'x_i\right) \\
\phi\left(\boldsymbol{w}_i,\bar{\boldsymbol{\theta}}\right) &\equiv E\left(x_i'x_i\right)^{-1}\left(x_i \cdot (y_i - x_i\beta)\right)
\end{aligned}$$

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: Bootstrap (bad finite-sample properties)
- Option 2: Just estimate them jointly in a big GMM system.
  - This option can be cumbersome if you have many moments.
- Option 3: Stack the influence functions and take the inner product.[2]
- 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.

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

Stacking

- Let's define

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

- Notice I dropped the i subscripts. What does this look like if there are 4 regressors?

	A	B	C	D	E
1	ϕ_μ	$\phi_{\beta 1}$	$\phi_{\beta 2}$	$\phi_{\beta 3}$	$\phi_{\beta 4}$
2	-0.3077937	-0.2881243	-0.1493863	-0.4944986	0.09112854
3	-0.118798	0.35481714	-0.4654646	0.4161474	-0.2297822
4	0.27052049	-0.0679981	-0.2685441	-0.0713374	-0.2145772
5	0.23215481	-0.4571358	-0.3404458	0.00301565	0.20407327
6	0.19164567	0.19907597	0.49640239	-0.4696586	0.02729855
7	-0.4222975	-0.2423789	-0.477901	-0.1465544	-0.2059395
8	-0.0261526	-0.2232136	-0.275062	0.43224294	0.08048878
9	-0.2780236	-0.0397985	0.4320227	-0.2087948	-0.3908644
10	-0.2101804	0.44238463	0.371486	-0.1105543	-0.1978471
11	-0.0332542	0.38275707	0.31075324	-0.2856035	0.40799314
12	-0.2346248	0.21748158	0.19450942	-0.1521142	-0.253551
13	0.4381279	-0.1185061	0.04483009	-0.0954184	-0.4959698
14	-0.4511052	0.09310356	-0.1128563	-0.1910718	0.12800501
15	0.08186761	-0.2591236	0.21970691	0.09055461	-0.3267252
16	-0.4999294	-0.058372	-0.2427571	0.02993619	0.29671955
17	-0.2666323	0.44120731	0.15008241	-0.0226203	0.47779026

Stacking

- Let's reiterate:

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

- 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 Julia Code

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

# OLS influence function
bhat = inv(x'*x)*x'*y;
uhat = y - x*bhat;
olsinflnc = (inv((x'*x)./n) * ((x.*uhat)'))';

#Big influence function
biginflnc = zeros(size(x,1),size(x,2)+1);
biginflnc[:,1] = meaninflnc;
biginflnc[:,2:size(x,2)+1] = olsinflnc;

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

Outline

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

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

  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*.

Outline

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.
- 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:

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

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.

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.