Lecture Notes on Extremum Estimation for Econometrics II

Tim Armstrong

last updated: March 21, 2021

- These notes cover a general class of estimators called "extremum estimators," "M-estimators," "GMM," "indirect inference," etc. They are based mostly on Newey and McFadden (1994).
- Let $Q_0(\theta)$ be a population objective function function from Θ to $\mathbb{R}$ that depends on the population distribution. We sample from this population and form an estimate $\hat{Q}_n(\theta)$. An extremum estimator $\hat{\theta}$ is an estimate of $\theta_0 = \arg\max Q_0(\theta)$ given by $\arg\max \hat{Q}_n(\theta)$.
- Examples:
  - OLS: $Q_0(\beta) = -E(y_i - x_i'\beta)^2$, $\hat{Q}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n (y_i - x_i'\beta)^2$.
  - Nonlinear least squares (NLS): Suppose $E(y_i|x_i) = h(x_i, \theta_0)$. Then $h(x) = h(x_i, \theta_0)$ minimizes $E[(y_i - h(x_i))^2]$ over all functions h (shown earlier in the semester), so θ minimizes $Q_0(\theta) = -E[y_i - h(x_i, \theta)]^2$. Sample analogue: $\hat{Q}_n(\theta) = -\frac{1}{n}\sum_{i=1}^n (y_i - h(x_i, \theta))^2$.
  - Maximum likelihood: $y_i \sim f(\cdot, \theta_0)$, then $Q_0(\theta) = E \log f(y_i, \theta)$ maximized at $\theta = \theta_0$ (covered in metrics I). Sample analogue is the log likelihood function $\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^n f(y_i, \theta)$.
  - GMM: $Eg(z_i, \theta_0) = 0$ where $g(z, \theta)$ is a known function of the data z_i and θ. Let W be a positive definite symmetric matrix. Then θ_0 minimizes $Q_0(\theta) = -[Eg(z_i, \theta)]'W[Eg(z_i, \theta)]$. Sample analogue is $\hat{Q}_n(\theta) = -[n^{-1}\sum_{i=1}^n g(z_i, \theta)]'W_n[n^{-1}\sum_{i=1}^n g(z_i, \theta)]$ for a sequence W_n with $W_n \xrightarrow{p} W$.
    - We can cast nonlinear IV as a GMM estimator: if $y_i = h(x_i, \theta_0) + e_i$ and $E(e_i z_i) = 0$, we can take $g(x_i, z_i, \theta) = z_i(y_i - h(x_i, \theta))$.

* If $Q_0(\theta) = Eq(z_i, \theta)$, then we can also use the first order conditions: θ_0 sets $E\frac{d}{d\theta}q(z_i, \theta) = 0$. This gives the GMM framework with $g(z, \theta) = \frac{d}{d\theta}q(z, \theta)$. Thus, we can cast OLS, NLS and ML as GMM estimators.

- Much of empirical work in economics can be put into the GMM framework.

0.1 Consistency (Section 2 in NM)

- Even if $\hat{Q}_n(\theta) \xrightarrow{p} Q_0(\theta)$, we need extra conditions to get convergence of argmax. Uniform convergence is important here.

- thm. Suppose that (i) $Q_0(\theta)$ is uniquely maximized at θ_0 (ii) Θ is compact (iii) $Q_0(\theta)$ is continuous and (iv) $\sup_{\theta\in\Theta} |\hat{Q}_n(\theta) - Q_0(\theta)|$. Then $\hat{\theta} \xrightarrow{p} \theta_0$.

  pf. Omitted (see Thm. 2.1 in Newey and McFadden 1994).

- Note:
  - When (iv) holds, we say that $\hat{Q}_n(\theta)$ converges uniformly in probability to $Q_0(\theta)$.
  - Typically we do not worry about compactness. Often, we can use other arguments to show that the argmax over all of $\mathbb{R}^k$ is taken on a compact set with probability approaching 1.

0.1.1 Consistency for GMM

- GMM setup: θ_0 satisfies $g_0(\theta_0) = 0$ where $g_0(\theta) = Eg(z_i, \theta)$. $\hat{\theta}$ minimizes $\hat{g}_n(\theta)' W_n \hat{g}_n(\theta)$ where $\hat{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n g(z_i, \theta)$ and $W_n \xrightarrow{p} W$ with W positive definite.

- Lemma: Suppose that

$$g_0(\theta) = 0 \iff \theta_0 = 0 \qquad (*)$$

  and W is positive definite. Then $Q_0(\theta) = -g_0(\theta)' W g_0(\theta)$ is uniquely maximized at θ_0 (i.e. (i) holds for $Q_0(\theta)$).

  If $g_0(\theta)$ is continuous, then $Q_0(\theta)$ is continuous ((iii) holds). If $\sup_{\theta\in\Theta} |\hat{g}_n(\theta) - g_0(\theta)| \xrightarrow{p} 0$ and $g_0(\theta)$ is bounded over Θ, then $\sup_{\theta\in\Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$ ((iv) holds).

- Note:

– The GMM identification condition (*) is difficult to check. It is good practice to at least show that it holds for some specification of the dgp, even if only through Monte Carlo simulation. For tests for identification in GMM, see Wright (2003).

– To show $\sup_{\theta\in\Theta} |\hat{g}_n(\theta) - g_0(\theta)| \xrightarrow{p} 0$ we need a uniform law of large numbers, which can be obtained using, e.g., continuity of $g(z_i, \theta)$ in θ and bounds on $g(z_i, \theta)$ (see Lemma 2.4 of Newey and McFadden 1994).

0.1.2 Maximum Likelihood

- ML setup: $Q_0(\theta) = E \log f(y_i, \theta)$.
- We verified condition (i) for this model in Metrics I (lecture 18).
- Lemma: $Q_0(\theta) = E \log f(y_i; \theta)$ is uniquely maximized at θ_0 iff. the following condition holds:

$$\text{for all } \theta \neq \theta_0,\ P_{\theta_0}(f(y_i, \theta) \neq f(y_i, \theta_0)) > 0.$$

pf.: See Metrics I notes, lecture 18.

0.2 Asymptotic distribution (Section 3 in NM)

- Idea: expand around the first order conditions. We saw this for maximum likelihood in Metrics I (lecture 18).
- Expand around FOCs:

$$\begin{aligned} 0 = \frac{d}{d\theta}\hat{Q}_n(\hat{\theta}) &= \frac{d}{d\theta}\hat{Q}_n(\theta_0) + \left[\frac{d}{d\theta}\hat{Q}_n(\hat{\theta}) - \frac{d}{d\theta}\hat{Q}_n(\theta_0)\right] \\ &\approx \frac{d}{d\theta}\hat{Q}_n(\theta_0) + \left[\frac{d}{d\theta d\theta'}\hat{Q}_n(\theta_0)\right](\hat{\theta} - \theta_0) \end{aligned}$$

where the remainder term is $\mathcal{O}_P((\hat{\theta} - \theta_0)^2)$ under regularity conditions.

Let $\hat{H}(\theta) = \frac{d}{d\theta d\theta'}\hat{Q}_n(\theta)$. Then rearranging gives

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -\hat{H}(\theta_0)^{-1}\left[\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)\right].$$

Suppose that

$$\hat{H}(\theta_0) \xrightarrow{p} H(\theta_0) \equiv \frac{d}{d\theta d\theta'} Q_0(\theta_0) \qquad (*)$$

and

$$\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0) \xrightarrow{d} N(0, \Sigma). \qquad (**)$$

- Proposition: Suppose that (*) and (**) hold with $H = H(\theta_0)$ invertible. Then, under additional regularity conditions,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, H^{-1}\Sigma[H^{-1}]')$$

- Typically, we have

$$\frac{d}{d\theta}\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m(z_i, \theta)$$

for some function m, so we get (**) from a CLT. In this case

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -H^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(z_i, \theta_0).$$

- def.: If there is a function $\psi(z_i)$ with $E\psi(z_i) = 0$ such that

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(z_i) + o_P(1),$$

we say that $\hat{\theta}$ is asymptotically linear with influence function ψ.

- For extremum estimators, influence function is $\psi(z_i) = -H^{-1}m(z_i, \theta_0)$

0.2.1 ML

- ML setup: $Q_0(\theta) = E_{\theta_0} \log f(y_i, \theta)$. We have

$$\begin{aligned}
&\frac{d}{d\theta} Q_0(\theta) = \frac{d}{d\theta} E_{\theta_0} \log f(y_i, \theta) \\
&= E_{\theta_0} \frac{d}{d\theta} \log f(y_i, \theta) && \text{under regularity conditions} \\
&= E_{\theta_0} \frac{\frac{d}{d\theta} f(y_i, \theta)}{f(y_i, \theta)}
\end{aligned}$$

  At θ_0, this is $\frac{d}{d\theta} Q_0(\theta_0) = E_{\theta_0} \frac{\frac{d}{d\theta} f(y_i, \theta_0)}{f(y_i, \theta_0)} = \int \frac{d}{d\theta} f(y, \theta_0)\, dy = \frac{d}{d\theta} \int f(y, \theta_0)\, dy = \frac{d}{d\theta} 1 = 0$ (under regularity conditions that allow switching $\int$ and $\frac{d}{d\theta}$). For the second derivative,

$$\begin{aligned}
&\frac{d}{d\theta d\theta'} Q_0(\theta) = E_{\theta_0} \frac{d}{d\theta d\theta'} \log f(y_i, \theta) && \text{under regularity conditions} \\
&= E_{\theta_0} \frac{d}{d\theta} \frac{\frac{d}{d\theta'} f(y_i, \theta)}{f(y_i, \theta)} \\
&= E_{\theta_0} \frac{f(y_i, \theta) \frac{d}{d\theta d\theta'} f(y_i, \theta) - [\frac{d}{d\theta} f(y_i, \theta)][\frac{d}{d\theta} f(y_i, \theta)]'}{f(y_i, \theta)^2}.
\end{aligned}$$

  At $\theta = \theta_0$, the first term is $E_{\theta_0} \frac{\frac{d}{d\theta d\theta'} f(y_i, \theta_0)}{f(y_i, \theta_0)} = \int \frac{d}{d\theta d\theta'} f(y, \theta_0)\, dy = \frac{d}{d\theta d\theta'} \int f(y, \theta_0)\, dy = \frac{d}{d\theta d\theta'} 1 = 0$ (under regularity conditions that allow switching $\int$ and $\frac{d}{d\theta}$).

- This gives us the <u>information matrix equality</u>:

$$\frac{d}{d\theta d\theta'} Q_0(\theta_0) = E_{\theta_0} \frac{d}{d\theta d\theta'} \log f(y_i, \theta_0) = -E_{\theta_0} [\frac{d}{d\theta} \log f(y_i, \theta_0)][\frac{d}{d\theta_0} \log f(y_i, \theta)]'$$

- We call $-E_{\theta_0} \frac{d}{d\theta d\theta'} \log f(y_i, \theta_0)$ the <u>information matrix</u> and denote it $\mathcal{I}(\theta_0)$.

- Thus, for ML, we get the influence function representation with $H = -\mathcal{I}(\theta_0)$ and $m(z_i, \theta_0) = \frac{d}{d\theta} \log f(y_i, \theta_0)$:

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_0) &= \mathcal{I}(\theta_0)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{d}{d\theta} \log f(y_i, \theta_0) + o_P(1) \\
&\xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1} E_{\theta_0} [\frac{d}{d\theta} \log f(y_i, \theta_0)][\frac{d}{d\theta} \log f(y_i, \theta_0)]' \mathcal{I}(\theta_0)^{-1}) \\
&= N(0, \mathcal{I}(\theta_0)^{-1})
\end{aligned}$$

where the last step follows by the information matrix inequality.

- Estimate $\mathcal{I}(\theta_0)^{-1}$ using $\mathcal{I}(\hat{\theta})^{-1}$ or $\left\{\frac{1}{n}\sum_{i=1}^n[\frac{d}{d\theta}\log f(y_i,\hat{\theta})][\frac{d}{d\theta}\log f(y_i,\hat{\theta})]'\right\}^{-1}$ or

$$\mathcal{I}(\hat{\theta})^{-1}\left\{\frac{1}{n}\sum_{i=1}^n[\frac{d}{d\theta}\log f(y_i,\hat{\theta})][\frac{d}{d\theta}\log f(y_i,\hat{\theta})]'\right\}\mathcal{I}(\hat{\theta})^{-1}.$$

- The latter estimator gives valid inference if the full parametric model is misspecified but the true parameter still maximizes the likelihood (e.g. OLS is ML with homoskedastic normal errors, but we can use the sandwich formula under heteroskedastic errors).

- Often, we observe $\{(x_i', y_i)'\}_{i=1}^n$, and we only model the distribution of $y_i|x_i$ and not the marginal distribution of x_i. E.g., in the homoskedastic normal regression model, we have $y_i|x_i \sim N(x_i'\theta, \sigma^2)$, but we don't make assumptions on the margial distribution of x_i. In these cases, we can do ML with the conditional distribution of $y_i|x_i$:

$$\hat{Q}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f(y_i|x_i,\theta).$$

  Wooldridge (2010) calls this conditional ML. The results above giving the asymptotic distribution and variance estimate go through with minor changes. See Wooldridge (2010).

- Mention asymptotic efficiency of ML, Cramer-Rao lower bound.

0.2.2 GMM

- GMM setup: $\underset{k\times 1}{\theta_0}$ satisfies $g_0(\theta_0) = 0$ where $\underbrace{g_0(\theta)}_{m\times 1} = Eg(z_i,\theta)$. $\hat{\theta}$ minimizes $\hat{g}_n(\theta)'W_n\hat{g}_n(\theta)$ where $\hat{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n g(z_i,\theta)$ and $W_n \xrightarrow{p} W$ with W positive definite.

- Let $\hat{Q}_n(\theta) = -\frac{1}{2}\hat{g}_n(\theta)'W_n\hat{g}_n(\theta)$ and $Q_0(\theta) = -\frac{1}{2}g_0(\theta)'Wg_0(\theta)$.

- To apply the general result, we need to find $H = \frac{d}{d\theta d\theta'}Q_0(\theta_0)$ and get a CLT for $\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)$.

- Let $\underbrace{G(\theta)}_{m\times k} = \frac{d}{d\theta'}g_0(\theta)$ and $\underbrace{\hat{G}(\theta)}_{m\times k} = \frac{d}{d\theta'}\hat{g}_n(\theta)$ (note that we are using row derivatives rather than column vector derivatives). Applying the chain rule to the quadratic form

$(\frac{d}{dx}u(x)'Wu(x) = 2\left[\frac{d}{dx'}u(x)\right]'Wu(x))$, we have

$$\begin{aligned}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0) &= -\sqrt{n}\hat{G}(\theta_0)'W_n\hat{g}_n(\theta_0) \\ &\xrightarrow{d} N(0, G'W\Omega W'G)\end{aligned}$$

where $G = G(\theta_0)$ and $\Omega = Eg(z_i,\theta_0)g(z_i,\theta_0)'$. This gives (**) with $\Sigma = G'W\Omega W'G$.

To derive H, first note that

$$\frac{d}{d\theta}Q_0(\theta) = -G(\theta)'Wg_0(\theta).$$

The pth column of H is given by the derivative of this expression with respect θ_p at θ_0, which is

$$-G(\theta_0)'W\frac{d}{d\theta_p}g_0(\theta_0) - \left[\frac{d}{d\theta_p}G(\theta_0)\right]'W\underbrace{g_0(\theta_0)}_{=0}.$$

Stacking these columns over $p = 1,\ldots,k$ gives

$$H(\theta_0) = \frac{d}{d\theta d\theta'}Q_0(\theta_0) = -G'WG.$$

Applying the general result gives ...

- Proposition: Under regularity conditions, the GMM estimator satisfies

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\theta_0) &= -(G'WG)^{-1}G'W\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g(z_i,\theta_0) + o_P(1) \\ &\xrightarrow{d} N(0,(G'WG)^{-1}G'W\Omega W'G(G'WG)^{-1}).\end{aligned}$$

- Asymptotic variance estimation: Let $\hat{G} = \hat{G}(\hat{\theta})$ and let $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n}g(z_i,\hat{\theta})g(z_i,\hat{\theta})'$. Estimate $V_{GMM} \equiv (G'WG)^{-1}G'W\Omega W'G(G'WG)^{-1}$ with $\hat{V}_{GMM} \equiv (\hat{G}'W_n\hat{G})^{-1}\hat{G}'W_n\hat{\Omega}W_n'\hat{G}(\hat{G}'W_n\hat{G})^{-1}$ Then, under regularity conditions, $\hat{\Omega} \xrightarrow{p} \Omega$ and $\hat{G} \xrightarrow{p} G$, so that $\hat{V}_{GMM} \xrightarrow{p} V_{GMM}$.
  - Note on details and regularity conditions: We can apply a LLN to $\hat{G}(\theta_0)$, so showing consistency amounts to showing that we can replace $\hat{\theta}$ with θ_0. A uniform LLN suffices for this. Similar comments apply to $\hat{\Omega}$.

- Exactly identified case: When $m = k$, G is a square matrix so that $(G'WG)^{-1}G'W = G^{-1}W^{-1}[G']^{-1}G'W = G^{-1}$. This gives $V_{GMM} = G^{-1}\Omega[G^{-1}]'$.

  - GMM as exactly identified GMM with subset of moments: For any matrix $\underset{k\times m}{A}$, we can us $\underbrace{\tilde{g}(z_i,\theta_0)}_{k\times 1} = Ag(z_i,\theta_0)$ and get exactly identified GMM. The derivative matrix is then $\tilde{G}(\theta) = AG(\theta)$ and the variance of the moments is $\tilde{\Omega} = A'Eg(z_i,\theta_0)g(z_i,\theta_0)'A'$ which gives the asymptotic variance as $\tilde{G}^{-1}\tilde{\Omega}[\tilde{G}^{-1}]' = (AG)^{-1}A\Omega A'[(AG)^{-1}]'$. Thus, GMM with weighting matrix W is equivalent to exactly identified GMM using $Ag(z_i,\theta)$ for any A such that $(G'WG)^{-1}G'W = (AG)^{-1}A$. In particular, we can take $A = G'W$ (indeed, note that the FOCs for the GMM estimator take this form with $\hat{G}(\theta_0)$ in place of G).

- Efficient GMM: V_{GMM} is minimized (in positive definite sense) by taking $W = \Omega^{-1}$ (same arguments as Gauss-Markov for generalized least squares). Asymptotic variance is then $(G'\Omega^{-1}G)^{-1}$. We can form an efficient estimate using a two-step efficient GMM estimator:

Step 1 Form a GMM estimator $\hat{\theta}_{\text{init}}$ using a (possibly inefficient) W_n.

Step 2 Estimate Ω with $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(z_i,\hat{\theta}_{\text{init}})g(z_i,\hat{\theta}_{\text{init}})'$.

Step 3 Form the efficient GMM estimator as the GMM estimator with weighting matrix $W_n = \hat{\Omega}^{-1}$.

- Other Efficient GMM estimators: iterate more times, continuous updating GMM, empirical likelihood, etc. See Hansen and NM for details. We will discuss an estimator that achieves the efficient GMM asymptotic variance and has computational advantages.

- Example (nonlinear IV): $y_i = h(x_i,\theta_0) + e_i$ and $E(e_i z_i) = 0$ fits into GMM framework with $g(x_i,z_i,\theta) = z_i(y_i - h(x_i,\theta))$. We have

$$\frac{d}{d\theta'}g(x_i,z_i,\theta) = -z_i\frac{d}{d\theta'}h(x_i,\theta)$$

  so that

$$G(\theta) = -Ez_i\frac{d}{d\theta'}h(x_i,\theta), \quad \hat{G}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} z_i\frac{d}{d\theta'}h(x_i,\theta).$$

We also have

$$\Omega = E[z_i z_i'(y_i - h(x_i, \theta_0))^2].$$

If, in addition to the above assumptions, we have homoskedasticity ($E(e_i|z_i) = 0$ and $E(e_i^2|z_i) = \sigma^2$), then $\Omega = \sigma^2 E[z_i z_i']$. Thus, under homoskedasticity, $W_n = \frac{1}{n}\sum_{i=1}^n z_i z_i'$ is an efficient weighting matrix. We can estimate θ using two-step efficient GMM with $W_n = \frac{1}{n}\sum_{i=1}^n z_i z_i'$ as the initial weighting matrix, and with $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^n z_i z_i'(y_i - h(x_i, \hat{\theta}))^2$ as the weighting matrix in the second stage. Asymptotic variance can be estimated by plugging in $\hat{G}(\hat{\theta})$ and $\hat{\Omega}^{-1}$ to variance formula.

0.2.3 One-step estimators (3.4 in NM)

- Often, we have an initial estimator $\hat{\theta}_{\text{init}}$ ($\bar{\theta}$ in NM's notation), but do not want to optimize another complicated nonlinear objective function.
- Define the one (Newton-Raphson) step estimator:

$$\tilde{\theta} = \hat{\theta}_{\text{init}} - \hat{H}_{\text{init}}^{-1} \frac{d}{d\theta} \hat{Q}_n(\hat{\theta}_{\text{init}})$$

  where $\hat{H}_{\text{init}}$ is an estimator for $H = \frac{d}{d\theta d\theta'} Q_0(\theta)$ such as $\frac{d}{d\theta d\theta'} \hat{Q}_n(\hat{\theta}_{\text{init}})$.

  - The one-step estimator is formed by taking a single step of the Newton-Raphson optimization algorithm.
  - At the rth step, the Newton-Raphson algorithm takes a quadratic approximation to the objective function at the current value (say, θ_r) and gives the argmax of the quadratic approximation to the objective function at θ_r as the input to the $(r+1)$th step (say, θ_{r+1}).

- Proposition: Suppose that $\hat{\theta}_{\text{init}}$ is $\sqrt{n}$-consistent (i.e. $\sqrt{n}(\hat{\theta}_{\text{init}} - \theta_0) = \mathcal{O}_P(1)$) and $\hat{H}_{\text{init}} \xrightarrow{p} H_0$. Then, under additional regularity conditions, $\tilde{\theta}$ has the same asymptotic distribution as $\hat{\theta} = \arg\max \hat{Q}_n(\theta_0)$:

$$\sqrt{n}(\tilde{\theta} - \theta_0) = \sqrt{n}(\hat{\theta} - \theta_0) + o_P(1) \xrightarrow{d} N(0, H^{-1}\Sigma[H^{-1}]').$$

sketch of proof: We have, under the conditions for asymptotic normality of $\hat{\theta}$,

$$\begin{aligned}\sqrt{n}(\tilde{\theta}-\theta_0) &= \sqrt{n}(\hat{\theta}_{\text{init}}-\theta_0) - \sqrt{n}\hat{H}_{\text{init}}^{-1}\frac{d}{d\theta}\hat{Q}_n(\hat{\theta}_{\text{init}}) \\ &= \sqrt{n}(\hat{\theta}_{\text{init}}-\theta_0) - \sqrt{n}\hat{H}_{\text{init}}^{-1}\left[\frac{d}{d\theta}\hat{Q}_n(\theta_0) + \hat{H}(\theta_0)(\hat{\theta}_{\text{init}}-\theta_0) + o_P(\hat{\theta}_{\text{init}}-\theta_0)\right] \\ &= \underbrace{(I-\hat{H}_{\text{init}}^{-1}\hat{H}(\theta_0))}_{\xrightarrow{p}0}\underbrace{\sqrt{n}(\hat{\theta}_{\text{init}}-\theta_0)}_{=\mathcal{O}_P(1)} - \sqrt{n}\hat{H}_{\text{init}}^{-1}\frac{d}{d\theta}\hat{Q}_n(\theta_0) + o_P(1)\end{aligned}$$

where the second equality uses a second order Taylor expansion of $\hat{Q}_n(\theta)$ at θ_0. Since $\sqrt{n}(\hat{\theta}-\theta_0) = -\hat{H}(\theta_0)^{-1}\left[\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)\right] + o_P(1)$, it follows that the difference between the above display and $\sqrt{n}(\hat{\theta}-\theta_0)$ is $(\hat{H}(\theta_0)^{-1}-\hat{H}_{\text{init}}^{-1})\sqrt{n}\hat{Q}_n(\theta_0)+o_P(1)$, which converges in probability to zero.

- Efficient GMM: We can take $\hat{\theta}_{\text{init}}$ to be the GMM estimate with an initial (possibly inefficient) weighting matrix. Then, take $\hat{Q}_n(\theta) = -\frac{1}{2}\hat{g}_n(\theta)'\hat{\Omega}^{-1}\hat{g}_n(\theta)$ with $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^n g(z_i,\hat{\theta}_{\text{init}})g(z_i,\hat{\theta}_{\text{init}})'$. We could take $\hat{H}_{\text{init}}$ to be $\frac{d}{d\theta d\theta'}\hat{Q}_n(\hat{\theta}_{\text{init}})$, but this is cumbersome, since it involves second derivatives of $\hat{g}_n(\theta)$. Rather, let us use $\hat{H}_{\text{init}} = \hat{G}(\hat{\theta}_{\text{init}})'\hat{\Omega}^{-1}\hat{G}(\hat{\theta}_{\text{init}})$. This gives

$$\tilde{\theta} = \hat{\theta}_{\text{init}} - (\hat{G}(\hat{\theta}_{\text{init}})'\hat{\Omega}^{-1}\hat{G}(\hat{\theta}_{\text{init}}))^{-1}\hat{G}(\hat{\theta}_{\text{init}})'\hat{\Omega}^{-1}\hat{g}_n(\hat{\theta}_{\text{init}}).$$

0.3 Heuristic explanation of connection between Gauss-Markov and efficient GMM using one-step estimators

- Suppose we want to base estimators on $\hat{\theta}_{\text{init}}$ and $\hat{g}_n(\hat{\theta}_{\text{init}})$. Consider an estimator of the form

$$\tilde{\theta}_{K_n} = \hat{\theta}_{\text{init}} - K_n\hat{g}_n(\hat{\theta}_{\text{init}})$$

for some $k \times m$ matrix K_n. This is a generalization of the two-step estimator given above, and can be thought of as using $K_n\hat{g}_n(\hat{\theta}_{\text{init}})$ to estimate $\theta_0 - \hat{\theta}_{\text{init}}$.

- Assuming a Taylor expansion holds so that $\hat{g}_n(\hat{\theta}_{\text{init}}) = \hat{G}(\theta_0)(\hat{\theta}_{\text{init}}-\theta_0)+o_P(1/\sqrt{n})$, we

will have

$$\begin{aligned}\sqrt{n}(\tilde{\theta}_{K_n} - \theta_0) &= \sqrt{n}(\hat{\theta}_{\text{init}} - \theta_0) - K_n\sqrt{n}\hat{g}_n(\hat{\theta}_{\text{init}}) \\ &= \sqrt{n}(\hat{\theta}_{\text{init}} - \theta_0) - K_n\sqrt{n}\left[\hat{G}(\theta_0)(\hat{\theta}_{\text{init}} - \theta_0) + \hat{g}_n(\theta_0)\right] + o_P(1) \\ &= (I - K_n\hat{G}(\theta_0))\sqrt{n}(\hat{\theta}_{\text{init}} - \theta_0) - K_n\sqrt{n}\hat{g}_n(\theta_0) + o_P(1).\end{aligned}$$

If we choose a sequence of matrices $K_n \xrightarrow{p} K$ such that

$$KG = I, \tag{1}$$

then we will have

$$\sqrt{n}(\tilde{\theta}_{K_n} - \theta_0) \xrightarrow{d} N(0, K'\Omega^{-1}K). \tag{2}$$

- Choosing K to minimize the asymptotic variance (2) subject to (1) is the same as the problem of minimizing the variance of the linear estimator KY in the linear model with design matrix $-G$ with Ω as the variance of the error matrix ($y = -G\theta + \varepsilon$ where $var(\varepsilon) = \Omega$).

- In particular, we are using the approximation

$$\sqrt{n}\hat{g}(\tilde{\theta}_{\text{init}}) = -G\sqrt{n}(\theta_0 - \hat{\theta}_{\text{init}}) + \sqrt{n}\hat{g}(\theta_0) + o_P(1),$$

  which, asymptotically, takes the form of a linear regression model $y = X\beta + \varepsilon$ (with m observations and k parameters) with $\sqrt{n}(\theta_0 - \hat{\theta}_{\text{init}})$ playing the role of β, $\sqrt{n}\hat{g}(\hat{\theta}_{\text{init}})$ playing the role of y, $-G$ playing the role of the design matrix X, and $\sqrt{n}\hat{g}(\theta_0)$ playing the role of the error term ε.

- Note also that $KG = I$ holds iff. $K = (G'WG)^{-1}G'W$ for some matrix W. To see this, note that, if $KG = I$, then $(G'(K'K)G)^{-1}G'(K'K) = K$, so $K = (G'WG)^{-1}G'W$ with $W = K'K$ (note that this matrix is not invertible). Conversely, if $K = (G'WG)^{-1}G'W$, then it can be seen by inspection that $KG = I$. Thus, the set of possible K matrices that satisfy $KG = I$ are precisely those that are one-step GMM estimators for some weighting matrix W.

0.4 Asymptotic efficiency (Section 5 in NM)

- Consider two estimators $\hat{\theta}_a$ and $\hat{\theta}_b$ with $var(\hat{\theta}_a) = V_a$ and $var(\hat{\theta}_b) = V_b$. Suppose that

$$cov(\hat{\theta}_a, \hat{\theta}_a - \hat{\theta}_b) = 0. \qquad (*)$$

  This condition can be rewritten $var(\hat{\theta}_a) = cov(\hat{\theta}_a, \hat{\theta}_b)$, which gives

$$\begin{aligned} 0 \leq var(\hat{\theta}_a - \hat{\theta}_b) &= var(\hat{\theta}_a) + var(\hat{\theta}_b) - cov(\hat{\theta}_a, \hat{\theta}_b) - cov(\hat{\theta}_b, \hat{\theta}_a) \\ &= var(\hat{\theta}_a) + var(\hat{\theta}_b) - 2var(\hat{\theta}_a) = var(\hat{\theta}_b) - var(\hat{\theta}_a) \end{aligned}$$

  (where the inequality is in the positive definite sense).

- Thus, when (*) holds and $\hat{\theta}_a$ and $\hat{\theta}_b$ are both unbiased, $\hat{\theta}_a$ is weakly more efficient than $\hat{\theta}_b$. Same holds for asymptotic efficiency and asymptotic unbiasedness.

- Now suppose that we have a class of estimators $\hat{\theta}(\tau)$ indexed by $\tau \in T$. If τ^* satisfies $cov(\hat{\theta}(\tau^*), \hat{\theta}(\tau^*) - \hat{\theta}(\tau)) = 0$ for all $\tau \in T$, then $\hat{\theta}(\tau^*)$ has the minimum variance in this class.

- Let us apply this principle to extremum estimators. Suppose that, for some $D(\tau)$ and $m(z_i, \tau)$, estimators in the class satisfy

$$\begin{aligned} &\sqrt{n}(\hat{\theta}(\tau) - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D(\tau)^{-1} m(z_i, \tau) + o_P(1) \xrightarrow{d} N(0, V(\tau)) \\ &\text{where } V(\tau) = D(\tau)^{-1} E[m(z_i, \tau) m(z_i, \tau)'][D(\tau)^{-1}]' \end{aligned}$$

- For two estimators corresponding to τ^* and τ, we have

$$\begin{aligned} &acov(\hat{\theta}(\tau^*), \hat{\theta}(\tau^*) - \hat{\theta}(\tau)) \\ &= D(\tau^*)^{-1} E m(z_i, \tau^*) m(z_i, \tau^*)' [D(\tau^*)^{-1}]' - D(\tau^*)^{-1} E m(z_i, \tau^*) m(z_i, \tau)' [D(\tau)^{-1}]' \end{aligned}$$

  (here, $acov(\hat{\theta}_a, \hat{\theta}_a - \hat{\theta}_b)$ denotes the covariance of the limiting distribution of $\sqrt{n}((\hat{\theta}_a - \theta_0)', (\hat{\theta}_a - \hat{\theta}_b)')'$).

- Suppose that τ^* satisfies

$$D(\tau) = Em(z_i, \tau)m(z_i, \tau^*)' \text{ for all } \tau \in T. \qquad (**)$$

Then this gives

$$\begin{aligned} &acov(\hat{\theta}(\tau^*), \hat{\theta}(\tau^*) - \hat{\theta}(\tau)) \\ &= D(\tau^*)^{-1}D(\tau^*)'[D(\tau^*)^{-1}]' - D(\tau^*)^{-1}D(\tau)'[D(\tau)^{-1}]' = 0. \end{aligned}$$

This gives...

- thm.: Suppose τ^* satisfies (**). Then $\hat{\theta}(\tau^*)$ is efficient over $\tau \in T$: $V(\tau^*) \leq V(\tau)$ all $\tau \in T$.

0.4.1 Efficiency of ML among GMM estimators (5.1 in NM)

- Consider ML setup $z_i \sim f(z, \theta)$. We will show that the ML estimator is efficient among GMM estimators using the above framework. Since overidentified GMM is asymptotically equivalent to exactly identified GMM with moments $G'Wg(z_i, \theta)$, we can restrict attention to exactly identified GMM.

- For exactly identified GMM with moment function $g(z_i, \theta)$, we have $\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} -G(\theta_0)^{-1}g(z_i, \theta_0) + o_P(1)$, so we get the general setup with $g(\cdot, \theta)$ (the moment function) playing the role of τ and $m(z_i, \tau) = g(z_i, \theta_0)$ and $D(\tau) = -G(\theta_0)$.

- ML is GMM with $g(z_i, \theta) = \frac{d}{d\theta} \log f(z_i, \theta)$, so for showing efficiency of ML within GMM estimators, the condition (**) reduces to

$$-G(\theta_0) = Eg(z_i, \theta_0) \left[\frac{d}{d\theta} \log f(z_i, \theta_0) \right]' \qquad (***)$$

- For $g(z_i, \theta)$ to give consistent estimate for all θ, we need

$$0 = E_\theta g(z, \theta) = \int g(z, \theta) f(z, \theta)\, dz \quad \text{all } \theta \in \Theta$$

- Differentiating this identity (and exchanging differentiation and integration) gives

$$0 = \frac{d}{d\theta'} \int g(z,\theta) f(z,\theta)\, dz = \underbrace{\int \left[\frac{d}{d\theta'} g(z,\theta)\right] f(z,\theta)\, dz}_{=G(\theta_0) \text{ when evaluated at } \theta_0} + \underbrace{\int g(z,\theta) \left[\frac{d}{d\theta'} f(z,\theta)\right] dz}_{=E_{\theta_0} g(z_i,\theta_0) \frac{d}{d\theta} \log f(z_i,\theta_0) \text{ when evaluated at } \theta_0}.$$

  Evaluating this at $\theta = \theta_0$ gives

$$0 = G(\theta_0) + E g(z_i,\theta_0) \left[\frac{d}{d\theta} \log f(z_i,\theta)\right]',$$

  which reduces to the condition (***) as required.

- Thus, ML minimizes the asymptotic variance among all exactly identified GMM estimators.

  - Of course, GMM with other moment conditions may be preferable if we don't know the full parametric model.

0.4.2 Optimal instruments (5.4 in NM)

- Suppose we have moment conditions $E[\underbrace{\rho(w_i,\theta)}_{s\times 1} | z_i] = 0$ where $\theta \in \mathbb{R}^q$. Then, for any $\underset{q\times s}{A(z)}$, we have $EA(z_i)\rho(w_i,\theta) = 0$. Thus, we can perform (exactly identified) GMM with $g(z_i,w_i,\theta) = A(z_i)\rho(w_i,\theta)$.

- By influence function formula for exactly identified GMM, we have

$$\sqrt{n}(\hat{\theta} - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} - \left[EA(z_i) \frac{d}{d\theta'} \rho(w_i,\theta_0)\right]^{-1} A(z_i)\rho(w_i,\theta_0) + o_P(1)$$

  Thus, we have the general framework with $A(\cdot)$ playing the role of τ and

$$D(\tau) = -EA(z_i) \frac{d}{d\theta'} \rho(w_i,\theta_0) = -EA(z_i)G(z_i)$$

  where $G(z_i) = E\left[\frac{d}{d\theta'}\rho(w_i,\theta_0)|z_i\right]$ and $m(z_i,\tau) = A(z_i)\rho(w_i,\theta_0)$. Asymptotic variance

is

$$[EA(z_i)G(z_i)]^{-1}\left\{EA(z_i)\rho(w_i,\theta_0)\rho(w_i,\theta_0)'[A(z_i)]'\right\}\left\{[EA(z_i)G(z_i)]^{-1}\right\}'$$
$$=[EA(z_i)G(z_i)]^{-1}\left\{EA(z_i)\Omega(z_i)[A(z_i)]'\right\}\left\{[EA(z_i)G(z_i)]^{-1}\right\}'$$

where $\Omega(z_i)=E[\rho(w_i,\theta_0)\rho(w_i,\theta_0)'|z_i]$.

- Let A^* be the optimal A (assuming it exists). The condition (**) reduces to

$$\text{for all } A(\cdot), \qquad -EA(z_i)G(z_i)=EA(z_i)\rho(w_i,\theta_0)\rho(w_i,\theta_0)'[A^*(z_i)]'$$
$$=EA(z_i)\Omega(z_i)[A^*(z_i)]'.$$

- This will hold if we set

$$A^*(z_i)=-G(z_i)'[\Omega(z_i)]^{-1}.$$

The asymptotic variance is then

$$[EA^*(z_i)G(z_i)]^{-1}\left\{EA^*(z_i)\Omega(z_i)[A^*(z_i)]'\right\}\left\{[EA^*(z_i)G(z_i)]^{-1}\right\}'$$
$$=\left\{EG(z_i)'[\Omega(z_i)]^{-1}G(z_i)\right\}^{-1}.$$

- Note that, for any $q\times q$ invertible constant matrix C, the GMM estimator is not changed if we replace $A^*(z_i)$ with $CA^*(z_i)$. For example, we can take $A^*(z_i)=G(z_i)'[\Omega(z_i)]^{-1}$ (i.e. we can remove the minus sign).

- Example: Consider linear IV where $y_i=x_i'\theta+e_i$ and we impose $E[e_i|z_i]=0$ (stronger than just imposing $Ez_ie_i=0$, as we did before). This fits into the framework with $\rho(y_i,x_i,\theta)=y_i-x_i'\theta$ and $s=1$. The GMM estimator with the moment function $\underbrace{A(z_i)}_{q\times 1}(y_i-x_i'\theta)$ is just the (exactly identified) 2SLS estimator with instruments $A(z_i)$.

Note that $G(z_i)=\frac{d}{d\theta'}E(y_i-x_i'\theta|z_i)=-E(x_i|z_i)'$ and $\Omega(z_i)=E[e_i^2|z_i]$, which gives the optimal instruments as $A^*(z_i)=E(x_i|z_i)/E[e_i^2|z_i]$. When e_i is homoskedastic ($E[e_i^2|z_i]=\sigma^2$ is constant) and $E(x_i|z_i)$ is linear, this gives the IV estimator with instruments $E(x_i|z_i)=\Gamma'z_i$ where $\Gamma=E(z_iz_i')E(z_ix_i')$, which is (using IV matrix

notation) $(\Gamma' Z' X)^{-1} \Gamma' Z' y$. Asymptotic variance is

$$\left\{ E\Gamma' z_i \sigma^{-2} z_i' \Gamma \right\}^{-1} = \sigma^2 \left\{ \Gamma' E z_i z_i' \Gamma \right\}^{-1} = \sigma^2 \left\{ E(z_i x_i')[E z_i z_i']^{-1} E z_i' x_i \right\}^{-1}.$$

Note that the 2SLS estimator is $(X' P_Z X)^{-1} X' P_Z y = (\hat{\Gamma}' Z' X)^{-1} \hat{\Gamma}' Z' y$. Thus, the 2SLS estimator can be interpreted as estimating the optimal instruments $\Gamma' z_i$ and applying just-identified GMM. Note that (under homoskedasticity), 2SLS has the same asymptotic variance even though the instruments are estimated:

$$\sqrt{n}(\hat{\theta}_{2SLS} - \theta_0) \xrightarrow{p} N(0, \sigma^2 \left\{ E(z_i x_i')' [E(z_i z_i')]^{-1} E(z_i x_i') \right\}^{-1}).$$

- Feasible plug-in estimators: If we can estimate $\Omega(z_i) = E[\rho(w_i, \theta)\rho(w_i, \theta)' | z_i]$ and $G(z_i) = E\left[\frac{d}{d\theta'} \rho(w_i, \theta) | z_i\right]$, then we can plug these in and use the optimal GMM estimator with $\hat{A}^*(z_i) = \hat{G}(z_i)' [\hat{\Omega}(z_i)]^{-1}$. As we saw above, 2SLS can be considered an example of this. More generally, we can estimate these objects using the theory of nonparametric function estimation, which we will cover later (time permitting). Often, we will be able to do this without affecting the asymptotic distribution (see discussion in NM Section 5.5).

- In practice, $G(z_i)$ and $\Omega(z_i)$ can be difficult to estimate, particularly if z_i is of moderate to high dimension. Economic models or intuition about the relation between z_i and w_i can be useful in forming estimates or guesses of the optimal instruments. One example of this is Berry et al. (1999) and, more recently, Reynaert and Verboven (2014).

0.5 Two-step plug-in estimators

- We will consider two-step estimators, in which we plug in an estimator $\hat{\gamma}$ in order to form an estimator $\hat{\theta}$.

- We will consider the case where each step is exactly identified GMM. Let $\hat{\theta}$ solve the equation

$$0 = \hat{g}_n(\theta, \hat{\gamma}) = \frac{1}{n} \sum_{i=1}^{n} g(z_i, \theta, \hat{\gamma})$$

where $\hat{\gamma}$ solves

$$0 = \hat{m}_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} m(z_i, \gamma).$$

- To find the asymptotic variance of θ, we can note that $(\hat{\theta}', \hat{\gamma}')$ is a GMM estimator based on the "stacked" sample equations

$$0 = \begin{pmatrix} \hat{g}_n(\theta, \gamma) \\ \hat{m}_n(\gamma) \end{pmatrix}$$

which estimate the stacked population equations

$$0 = \begin{pmatrix} g_0(\theta_0, \gamma_0) \\ m_0(\gamma_0) \end{pmatrix} = \begin{pmatrix} Eg(z_i, \theta_0, \gamma_0) \\ Em(z_i, \gamma_0) \end{pmatrix}.$$

- Asymptotic variance of $(\hat{\theta}', \hat{\gamma}')'$ is $G_{\text{stacked}}^{-1}\Omega_{\text{stacked}}[G_{\text{stacked}}^{-1}]'$ where

$$\begin{aligned} G_{\text{stacked}} &= \frac{d}{d(\theta', \gamma')} \begin{pmatrix} Eg(z_i, \theta_0, \gamma_0) \\ Em(z_i, \gamma_0) \end{pmatrix} = \begin{pmatrix} \frac{d}{d\theta'} Eg(z_i, \theta_0, \gamma_0) & \frac{d}{d\gamma'} Eg(z_i, \theta_0, \gamma_0) \\ 0 & \frac{d}{d\gamma'} Em(z_i, \gamma_0) \end{pmatrix} \\ &\equiv \begin{pmatrix} G_\theta & G_\gamma \\ 0 & M \end{pmatrix} \end{aligned}$$

and

$$\Omega_{\text{stacked}} = \begin{pmatrix} Eg(z_i, \theta_0, \gamma_0)g(z_i, \theta_0, \gamma_0)' & Eg(z_i, \theta_0, \gamma_0)m(z_i, \gamma_0)' \\ Em(z_i, \gamma_0)g(z_i, \theta_0, \gamma_0)' & Em(z_i, \gamma_0)m(z_i, \gamma_0)' \end{pmatrix}$$

- Using the partitioned matrix inverse formula, we have

$$G_{\text{stacked}}^{-1} = \begin{pmatrix} G_\theta^{-1} & -G_\theta^{-1}G_\gamma M^{-1} \\ 0 & M^{-1} \end{pmatrix}$$

which gives the asymptotic variance as

$$E \begin{pmatrix} G_\theta^{-1} & -G_\theta^{-1}G_\gamma M^{-1} \\ 0 & M^{-1} \end{pmatrix} \begin{pmatrix} g(z_i, \theta_0, \gamma_0) \\ m(z_i, \gamma_0) \end{pmatrix} \begin{pmatrix} g(z_i, \theta_0, \gamma_0) \\ m(z_i, \gamma_0) \end{pmatrix}' \begin{pmatrix} G_\theta^{-1} & -G_\theta^{-1}G_\gamma M^{-1} \\ 0 & M^{-1} \end{pmatrix}'.$$

The asymptotic variance of $\hat{\theta}$ is the top-left element:

$$\sqrt{n}(\hat{\theta} - \theta_0)$$
$$\xrightarrow{d} N(0, G_\theta^{-1} E[g(z_i, \theta_0, \gamma_0) - G_\gamma M^{-1} m(z_i, \gamma_0)][g(z_i, \theta_0, \gamma_0) - G_\gamma M^{-1} m(z_i, \gamma_0)]'[G_\theta^{-1}]').$$

- Estimating the asymptotic variance: we can use the above formula with G_θ replaced by $\frac{1}{n}\sum_{i=1}^n \frac{d}{d\theta'} g(z_i, \hat{\theta}, \hat{\gamma})$, etc. This corresponds to the upper-left portion of the estimate of the asymptotic variance matrix of $(\hat{\theta}', \hat{\gamma}')'$ estimated using the usual formula. This gives a straightforward way of getting variance estimates for two-step estimates:

  1.) Write $\hat{\gamma}$ as a GMM estimator with moment function $m(z_i, \gamma)$ and $\hat{\theta}$ as a GMM estimator with moment function $g(z_i, \theta, \hat{\gamma})$.

  2.) Stack the two moment functions into the moment function $g_{\text{stacked}}(z, \theta, \gamma) = (g(z_i, \theta, \gamma)', m(z_i, \gamma)')'$.

  3.) Compute the asymptotic variance estimate for the stacked model, and take the variance estimate for $\hat{\theta}$ as the part of that variance matrix corresponding to θ.

  For many moment functions g that are composed of simple functions such as polynomials, logs, etc., this can now be done using STATA.

- When does using $\hat{\gamma}$ rather than γ_0 affect the asymptotic variance? Note that the asymptotic variance with $\hat{\gamma}$ replaced by γ_0 is $G_\theta^{-1} E g(z_i, \theta_0, \gamma_0) g(z_i, \theta_0, \gamma_0)'[G_\theta^{-1}]'$. Thus, the two asymptotic variances coincide when $G_\gamma = 0$. This holds when $Eg(z_i, \theta_0, \gamma) = 0$ holds for all γ, not just γ_0. In other words, this holds when the moment function for θ is "correctly specified" regardless of γ.

- Example: Linear conditional mean model

$$y_i = x_i'\theta + e_i, \quad E(e_i|x_i) = 0.$$

  Let $\sigma^2(x_i) = E(e_i^2|x_i)$. The weighted least squares (WLS) estimator is the GMM estimator with moments

$$x_i(y_i - x_i'\theta)\frac{1}{\sigma^2(x_i)}.$$

  It is a special case of generalized least squares (see 4.7 and 12.2 in 2017 version of

Hansen's text), and corresponds to the optimal instruments in the linear IV example above when $z_i = x_i$.

Suppose we have a parametric model for $\sigma^2(x_i)$, say,

$$\sigma^2(x_i) = \sigma^2(x_i, \gamma_1) = x_i'\gamma_1.$$

This suggests applying WLS based on an estimated γ_1. Consider the following procedure:

step 1: Regress y_i on x_i. Call the residuals $\hat{u}_i$.

step 2: Regress $\hat{u}_i^2$ on x_i to get the estimate $\hat{\gamma}_1$

step 3: Perform the weighted least squares regression of y_i on x_i with weights $1/\sigma^2(x_i, \hat{\gamma}_1)$.

To phrase this procedure as a single set of stacked GMM moments, we can apply the general framework with

$$m(x_i, y_i, \gamma_1, \gamma_2) = \begin{pmatrix} x_i\left[(y_i - x_i'\gamma_2)^2 - x_i'\gamma_1\right] \\ x_i\left(y_i - x_i'\gamma_2\right) \end{pmatrix}$$

and

$$g(z_i, \theta, \gamma_1, \gamma_2) = x_i(y_i - x_i'\theta)\frac{1}{\sigma^2(x_i, \tilde{\gamma}_1)} = 0$$

(here γ_2 is equal to θ, but we allow the estimate used in the first stage to differ from the final estimate for θ). Since $Eg(z_i, \theta, \tilde{\gamma}_1, \tilde{\gamma}_2) = Ex_i(y_i - x_i'\theta)\frac{1}{\sigma^2(x_i, \tilde{\gamma}_1)} = 0$ even for $\tilde{\gamma}_1$ not equal to the true γ_1, we will have $G_\gamma = 0$ so that estimating γ does not affect the asymptotic distribution of $\hat{\theta}$.

  - In general, we can estimate $A(z)$ when using the moments $EA(z_i)\rho(w_i, \theta) = 0$ whenever we have the model $E[\rho(w_i, \theta)|z_i] = 0$.

- See NM for example of Heckman two-step estimator for selection model (running example in Section 6).

- Derivation based on influence functions: Note that $\psi(z_i) \equiv M^{-1}m(z_i, \gamma_0)$ is the influence funtion for $\hat{\gamma}$. Thus, the asymptotic variance of $\hat{\theta}$ can be written

$$G_\theta^{-1}E[g(z_i, \theta_0, \gamma_0) - G_\gamma\psi(z_i)][g(z_i, \theta_0, \gamma_0) - G_\gamma\psi(z_i)]'[G_\theta^{-1}]'.$$

This holds more generally whenever $\hat{\gamma}$ has an influence function representation. This follows since

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_0) &= -G_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0, \hat{\gamma}) + o_P(1) \\
&= -G_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0, \gamma_0) - G_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n [g(z_i, \theta_0, \hat{\gamma}) - g(z_i, \theta_0, \gamma_0)] + o_P(1) \\
&= -G_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0, \gamma_0) - \sqrt{n} G_\theta^{-1} \hat{G}_\gamma(\gamma_0)(\hat{\gamma} - \gamma_0) + o_P(1) \\
&= -G_\theta^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0, \gamma_0) - G_\theta^{-1} G_\gamma \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(z_i) + o_P(1)
\end{aligned}$$

where the third equality uses a Taylor approximation to $\frac{1}{n}\sum_{i=1}^n g(z_i, \theta_0, \gamma)$ and the last equality uses the influence function representation for γ.

Proposition: Let $\hat{\theta}$ solve the moment conditions

$$0 = \frac{1}{n} \sum_{i=1}^n g(z_i, \theta, \hat{\gamma})$$

(with $\hat{\gamma}$ plugged in). Suppose that

$$\sqrt{n} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n g(z_i, \theta_0, \gamma_0) \\ \hat{\gamma} - \gamma_0 \end{pmatrix} \xrightarrow{d} \begin{pmatrix} Z_g \\ Z_\gamma \end{pmatrix}$$

for some limiting random variables Z_g, Z_γ. Then, under regularity conditions

$$\begin{aligned}
\sqrt{n}(\hat{\theta} - \theta_0) &= -G_\theta^{-1} \left[\frac{1}{\sqrt{n}} \sum_{i=1}^n g(z_i, \theta_0, \gamma_0) + G_\gamma \sqrt{n}(\hat{\gamma} - \gamma_0) \right] + o_P(1) \\
&\xrightarrow{d} -G_\theta^{-1} \left(Z_g + G_\gamma Z_\gamma \right).
\end{aligned}$$

In particular, if

$$\begin{pmatrix} Z_g \\ Z_\gamma \end{pmatrix} \sim N\left(0, \begin{pmatrix} \Omega_g & \Omega_{g\gamma} \\ \Omega_{\gamma g} & \Omega_\gamma \end{pmatrix}\right),$$

then

$$\sqrt{n}(\hat{\theta}-\theta_0) \xrightarrow{d} G_\theta^{-1} \begin{pmatrix} I & G_\gamma \end{pmatrix} \begin{pmatrix} \Omega_g & \Omega_{g\gamma} \\ \Omega_{\gamma g} & \Omega_\gamma \end{pmatrix} \begin{pmatrix} I \\ G_\gamma' \end{pmatrix} [G_\theta^{-1}]'$$

- Sometimes $\hat{\gamma}$ is obtained from a data set that is drawn independently of the z_i's. In this case, $\Omega_{g\gamma} = \Omega_{\gamma g}' = 0$.

0.6 Nonsmooth objective functions

- See separate notes on quantile regression/nonsmooth objective function.

(skipping section 8 from NM on semiparametric two-step estimators)

0.7 Hypothesis testing (Section 9 in NM)

- Extremum estimator setup: $\underset{k\times 1}{\hat{\theta}}$ maximizes $\hat{Q}_n(\theta)$ where $\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0) \xrightarrow{d} N(0,\Sigma)$ and $\frac{d}{d\theta d\theta'}\hat{Q}_n(\theta_0) \xrightarrow{p} \frac{d}{d\theta d\theta'}Q_0(\theta_0) \equiv H$.
  - Under regularity conditions, we showed $\sqrt{n}(\hat{\theta}-\theta_0) = -H^{-1}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)+o_P(1) \xrightarrow{d} N(0, H^{-1}\Sigma H^{-1})$.
  - When $-H = \Sigma$ (efficient GMM, correctly specified ML), this simplified to $N(0,\Sigma^{-1})$. We will see that this condition simplifies certain hypothesis tests.
  - GMM is a special case: $\hat{Q}_n(\theta) = -\frac{1}{2}\hat{g}_n(\theta)'W\hat{g}_n(\theta)$, $\Sigma = G'W\Omega^{-1}WG$ and $H = -G'WG$. We get $-H = \Sigma$ when $W = -\Omega^{-1}$ (efficient GMM).
- Null hypothesis $H_0 : \underbrace{a(\theta)}_{r\times 1} = 0$, $r \leq k$. Let $\underset{r\times k}{A} = \frac{d}{d\theta'}a(\theta)$. Assume A is full rank.
- Define the <u>constrained estimator</u>

$$\bar{\theta} = \arg\max_\theta \hat{Q}_n(\theta) \quad \text{s.t.} \quad a(\theta) = 0.$$

- Lagrangian:

$$\mathcal{L} = \hat{Q}_n(\theta) - a(\theta)'\gamma$$

- Constrained estimator $\bar{\theta}$ satisfies FOCs:

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{d}{d\theta}\hat{Q}_n(\bar{\theta}) - \frac{d}{d\theta}a(\bar{\theta})'\bar{\gamma} \\ -a(\bar{\theta}) \end{pmatrix}$$

  where $\bar{\gamma} = \bar{\gamma}_n$ is the vector of Lagrange multipliers at the maximum.

- We will consider the "trinity" of test statistics. Suppose $H = -\Sigma$. Let $\hat{\Sigma}$ be a consistent estimate of Σ (since $H = -\Sigma$, we can use $\hat{\Sigma} = -\hat{H} = -\frac{d}{d\theta d\theta'}\hat{Q}_n(\hat{\theta})$ or $\hat{\Sigma} = -\hat{H} = -\frac{d}{d\theta d\theta'}\hat{Q}_n(\bar{\theta})$).

  - Wald statistic: based on $a(\hat{\theta}) - a(\bar{\theta})$ or $\hat{\theta} - \bar{\theta}$. When $H = -\Sigma$, we can use

$$W = na(\hat{\theta})'[\hat{A}\hat{\Sigma}^{-1}\hat{A}']^{-1}a(\hat{\theta})$$

    (we showed this earlier).

  - Lagrange multiplier (LM) or score statistic: based on $\bar{\gamma}$ or $\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})$. When $H = -\Sigma$, we can use

$$LM = n\left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right]'\hat{\Sigma}^{-1}\left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right].$$

    * Advantage of LM statistic: if we use a version of $\hat{\Sigma}$ based on $\bar{\theta}$, we do not need to find $\hat{\theta}$ (unconstrained estimate) to compute it.

  - Distance metric (DM) statistic: based on $\hat{Q}_n(\hat{\theta}) - \hat{Q}_n(\bar{\theta})$. We need to have $H = -\Sigma$ to use this statistic. It is defined as

$$DM = 2n[\hat{Q}_n(\hat{\theta}) - \hat{Q}_n(\bar{\theta})].$$

  - Advantage of DM statistic: don't need to compute derivatives of $\hat{Q}_n(\theta)$.

  (Note: these test statistics are defined on p.2226 of NM. NM provide the derivation for the GMM case - see Section 7.4 in Hayashi for a derivation along the lines of the one given here.)

- Recall that, in the linear regression model, the Wald test could also be interpreted as a LM or DM test (we showed that it is equivalent to rejecting for large values of the Lagrange multiplier or DM). More generally, this is the case when $\hat{Q}_n$ is quadratic, which holds in GMM when $\hat{g}_n(\theta)$ is linear.

- Proposition: Suppose that $H = -\Sigma$ and additional regularity conditions hold. Then $W \xrightarrow{d} \chi^2_r$, $LM \xrightarrow{d} \chi^2_r$ and $DM \xrightarrow{d} \chi^2_r$.

- We covered the Wald statistic based on $a(\hat{\theta})$ earlier in the general setting of asymptotically normal estimators (just use delta method and asymptotic distribution of $\hat{\theta}$). Thus, we will focus on LM and DM statistics and give a heuristic derivation.

- Under regularity conditions, we have

$$\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\bar{\theta}) = \sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0) + \underbrace{\sqrt{n}\frac{d}{d\theta d\theta'}\hat{Q}_n(\theta_0)}_{\xrightarrow{p} H}(\bar{\theta}-\theta_0) + \underbrace{\mathcal{O}_P(\sqrt{n}(\bar{\theta}-\theta_0)^2)}_{=o_P(1)}$$

  and

$$\sqrt{n}a(\bar{\theta}) = \sqrt{n}a(\theta_0) + \sqrt{n}A(\bar{\theta}-\theta_0) + \underbrace{\mathcal{O}_P(\sqrt{n}(\bar{\theta}-\theta_0)^2)}_{=o_P(1)}.$$

  Also, assuming $a(\theta)$ is continuously differentiable, we will have

$$\frac{d}{d\theta'}a(\bar{\theta}) \xrightarrow{p} A.$$

- Scaling FOCs by $\sqrt{n}$ and plugging this in gives

$$\begin{aligned}
\begin{pmatrix} 0 \\ 0 \end{pmatrix} &= \begin{pmatrix} \sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0) + \sqrt{n}H(\bar{\theta}-\theta_0) - \sqrt{n}A'\bar{\gamma} \\ -\sqrt{n}A(\bar{\theta}-\theta_0) \end{pmatrix} + o_P(1) \\
\implies \begin{pmatrix} \sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0) \\ 0 \end{pmatrix} &= \begin{pmatrix} -\sqrt{n}H(\bar{\theta}-\theta_0) + \sqrt{n}A'\bar{\gamma} \\ \sqrt{n}A(\bar{\theta}-\theta_0) \end{pmatrix} + o_P(1) \\
&= \sqrt{n}\begin{pmatrix} -H & A' \\ A & 0 \end{pmatrix}\begin{pmatrix} \bar{\theta}-\theta_0 \\ \bar{\gamma} \end{pmatrix} + o_P(1).
\end{aligned}$$

  By the partitioned matrix inverse formula, we have

$$\begin{pmatrix} -H & A' \\ A & 0 \end{pmatrix}^{-1} = \begin{pmatrix} -H^{-1} + H^{-1}A'(A'H^{-1}A)^{-1}AH^{-1} & H^{-1}A'(AH^{-1}A')^{-1} \\ (AH^{-1}A')^{-1}AH^{-1} & (A'H^{-1}A)^{-1} \end{pmatrix}$$

- Putting this together gives

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\bar{\theta}-\theta_0\\ \bar{\gamma}\end{pmatrix} &= \begin{pmatrix}-H & A'\\ A & 0\end{pmatrix}^{-1}\begin{pmatrix}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)\\ 0\end{pmatrix}+o_P(1)\\ &= \begin{pmatrix}-(H^{-1}-H^{-1}A'(A'H^{-1}A)^{-1}AH^{-1})\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)\\ (A'H^{-1}A)^{-1}AH^{-1}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)\end{pmatrix}+o_P(1).\end{aligned}$$

  Since $\sqrt{n}(\hat{\theta}-\theta_0) = -H^{-1}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)+o_P(1)$, this gives

$$\begin{aligned}\sqrt{n}(\hat{\theta}-\bar{\theta}) &= -H^{-1}A'(A'H^{-1}A)^{-1}AH^{-1}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)+o_P(1)\\ &= -H^{-1}A'\sqrt{n}\bar{\gamma}+o_P(1).\end{aligned}$$

  Note also that, by FOCs,

$$\begin{aligned}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\bar{\theta}) &= \sqrt{n}\underbrace{\frac{d}{d\theta}a(\bar{\theta})'}_{\xrightarrow{p}A'}\bar{\gamma} = \sqrt{n}A'\bar{\gamma}+o_P(1)\\ &= A'(A'H^{-1}A)^{-1}AH^{-1}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)+o_P(1)\end{aligned}$$

- Let $Z_n = \Sigma^{-1/2}\sqrt{n}\frac{d}{d\theta}\hat{Q}_n(\theta_0)$ so that $Z_n \xrightarrow{d} Z \sim N(0, I_k)$. This gives

$$\begin{aligned}LM &= n\left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right]'\hat{\Sigma}^{-1}\left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right]\\ &= Z_n'[\Sigma^{1/2}]'H^{-1}A'(A'H^{-1}A)^{-1}A\hat{\Sigma}^{-1}A'(A'H^{-1}A)^{-1}AH^{-1}\Sigma^{1/2}Z_n+o_P(1)\\ &\xrightarrow{d} Z'[\Sigma^{1/2}]'H^{-1}A'(A'H^{-1}A)^{-1}A\Sigma^{-1}A'(A'H^{-1}A)^{-1}AH^{-1}\Sigma^{1/2}Z.\end{aligned}$$

  Since $H = -\Sigma$, this simplifies to

$$Z'\Sigma^{-1/2}A'(A'\Sigma^{-1}A)^{-1}A[\Sigma^{-1/2}]'Z.$$

  The matrix in the middle is symmetric and idempotent with rank r (same rank as A), so this is the χ^2_r distribution.

- For the DM statistic, note that, under regularity conditions, a second order expansion

of $\hat{Q}_n(\theta)$ around its maximum $\hat{\theta}$ gives

$$DM = 2n[\hat{Q}_n(\hat{\theta}) - \hat{Q}_n(\bar{\theta})] = -2n \cdot \frac{1}{2}(\hat{\theta} - \bar{\theta})' \underbrace{\frac{d}{d\theta d\theta'}\hat{Q}_n(\theta_0)}_{\xrightarrow{p} H}(\hat{\theta} - \bar{\theta}) + o_P(1).$$

By the derivations above, $\sqrt{n}(\hat{\theta} - \bar{\theta}) = -\sqrt{n}H^{-1}A'\bar{\gamma} + o_P(1) = -\sqrt{n}H^{-1}\frac{d}{d\theta}\hat{Q}_n(\bar{\theta}) + o_P(1)$. Combining this with the above display gives

$$DM = \left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right]' H^{-1}HH^{-1}\left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right] + o_P(1) = \left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right]' \Sigma^{-1}\left[\frac{d}{d\theta}\hat{Q}_n(\bar{\theta})\right] + o_P(1)$$

where the last equality uses $H = -\Sigma$. Thus, $DM \xrightarrow{d} \chi^2_r$ by the derivations for LM.

- Maximum likelihood: $\Sigma = -H$ holds by the information matrix equality. The DM test is called the likelihood ratio test, and the LM/score test is called the score test.
- GMM: We get $\Sigma = -H$ when $W = \hat{\Omega}^{-1} \xrightarrow{p} \Omega^{-1}$ (note that multiplying the objective function by $1/2$ is crucial here). Note that we can use the general version of the Wald test derived earlier when W does not converge to Ω^{-1}.
- Asymptotic equivalence: The trinity of tests are "asymptotically equivalent" in the sense that the local asymptotic power functions are identical for alternatives of the form $a(\theta) = \delta/\sqrt{n}$ (see NM; see Metrics I for definition of local asymptotic power function).

0.7.1 Constrained estimation

- If we are willing to impose the condition $a(\theta_0) = 0$, then it may make sense to report the constrained estimator $\bar{\theta}$. The asymptotic variance of $\bar{\theta}$ can be obtained using the asymptotic approximation for $\sqrt{n}(\bar{\theta} - \theta_0)$ above. This can be used to form standard errors. See NM for derivation in the case of efficient GMM.

0.7.2 Test of Overidentifying Restrictions

- GMM with $\underbrace{g(\theta)}_{m\times 1}$, $\theta \in \mathbb{R}^k$. If $m > k$, there are testable restrictions: if there does not exist θ such that $g(\theta) = 0$, then the model is misspecified.

- We will consider a test of the null that the model is correctly specified:

$$H_0 : \text{there exists } \theta_0 \text{ s.t. } g(\theta_0) = 0.$$

- The J-statistic is defined as the minimized (scaled) GMM objective function with optimal weighting:

$$J = -2n\hat{Q}_n(\hat{\theta}) = n[\hat{g}(\hat{\theta})]'\hat{\Omega}^{-1}[\hat{g}(\hat{\theta})]$$

where $\hat{\theta}$ is the GMM estimate with $W = \hat{\Omega}^{-1}$ and $\hat{\Omega}$ is a consistent estimate of Ω.

  - Sometimes called "Sargan J-statistic" or "Hansen J-statistic."

- Proposition: Suppose that the model is correctly specified and additional regularity conditions hold. Then

$$J \xrightarrow{d} \chi^2_{m-k}.$$

sketch of proof: Consider the GMM model with parameters $(\theta, \psi)'$ and moments $(g_1(\theta)', g_2(\theta)' - \psi')'$, where $g(\theta) = (g_1(\theta)', g_2(\theta)')'$ and g_1 contains the first k elements of g. If the original model is correctly specified, then $g(\theta_0) = 0$ so that $\psi = 0$. The DM statistic $H_0 : \psi = 0$ takes the form

$$2n[\tilde{Q}_n(\tilde{\theta}, \tilde{\psi}) - \tilde{Q}_n(\hat{\theta}, 0)]$$

where $\tilde{Q}_n(\theta, \psi) = -(1/2)(g_1(\theta)', g_2(\theta)' - \psi')\hat{\Omega}^{-1}(g_1(\theta)', g_2(\theta)' - \psi')'$ and $(\tilde{\theta}', \tilde{\psi}')'$ are the unconstrained estimates. Since the unconstrained model is exactly identified, $\tilde{Q}_n(\tilde{\theta}, \tilde{\psi}) = 0$. Combining this with the fact that $\tilde{Q}_n(\theta, 0) = \hat{Q}_n(\theta)$ gives the result.

0.7.3 Testing Overidentifying Restrictions in Linear IV

- Linear IV model: $y_i = x_i'\beta + e_i$, $Ez_ie_i = 0$, $\beta \in \mathbb{R}^k$, $z \in \mathbb{R}^\ell$.

- Suppose that $z = (z_1', z_2')'$ where z_2 is $r \times 1$. We wish to test the null that z is exogenous:

$$H_0 : \text{ there exists } \beta \text{ s.t. } Ez_i(y_i - x_i'\beta) = 0$$

against the alternative that z_1 is exogenous, but z_2 is not:

$$H_1 : Ez_{1,i}(y_i - x_i'\beta) = 0 \text{ but } Ez_{2,i}(y_i - x_i'\beta) \neq 0.$$

Note: in order to "conclude" that z_2 is endogenous when we reject H_0, we are maintaining the hypothesis that z_1 is exogenous. Alternatively, we can just test H_0 and think of it as a general specification test: if we reject, it just means that at least one element of z is endogenous.

- We can test this by considering the larger model $y_i = x_i'\beta + z_{2,i}'\psi + e_i$ with $Ez_ie_i = 0$. Then H_0 corresponds to the null hypothesis that $\psi = 0$ in this model.
- We can test it using Wald, LM or DM by placing this model in the general GMM framework. For DM and the version of LM we have covered, we need to use two-step GMM with an efficient weighting matrix.
- When $r = \ell - k$, this corresponds to a J-test.

0.8 Durbin-Wu-Hausman Tests

- Recall setup of Section 0.4: we had a class of estimators $\hat{\theta}(\tau)$ indexed by τ. We showed that, if τ^* satisfies

$$acov(\hat{\theta}(\tau^*), \hat{\theta}(\tau^*) - \hat{\theta}(\tau)) = 0 \quad \text{for all} \quad \tau, \qquad (*)$$

then $\hat{\theta}(\tau^*)$ is asymptotically efficient in this class.

- Suppose that $\hat{\theta}(\tau^*)$ uses additional constraints to gain efficiency. We want to know whether these constraints are true. Let $\hat{\theta}(\tau')$ be an estimator in the same class that does not use these constraints.
- We wish to test:

$$H_0 : \text{both } \hat{\theta}(\tau^*) \text{ and } \hat{\theta}(\tau') \text{ are consistent and } \tau^* \text{ satisfies } (*)$$

vs

$$H_1 : \hat{\theta}(\tau') \text{ is consistent, but } \hat{\theta}(\tau^*) \text{ is not}$$

- The condition (*) can be written $avar(\hat{\theta}(\tau^*)) = acov(\hat{\theta}(\tau^*), \hat{\theta}(\tau'))$, which gives

$$\begin{aligned} &avar(\hat{\theta}(\tau^*) - \hat{\theta}(\tau')) \\ &= avar(\hat{\theta}(\tau^*)) + avar(\hat{\theta}(\tau')) - acov(\hat{\theta}(\tau^*), \hat{\theta}(\tau')) - acov(\hat{\theta}(\tau'), \hat{\theta}(\tau^*)) \\ &= avar(\hat{\theta}(\tau')) - avar(\hat{\theta}(\tau^*)) \end{aligned}$$

- Thus, letting $V(\tau^*) = avar(\hat{\theta}(\tau^*))$ and $V(\tau') = avar(\hat{\theta}(\tau'))$ (so that $\sqrt{n}(\hat{\theta}(\tau^*) - \theta_0) \xrightarrow{d} N(0, V(\tau^*))$ under H_0 and similarly for $\hat{\theta}(\tau')$), we have

$$\sqrt{n}(\hat{\theta}(\tau^*) - \hat{\theta}(\tau')) \xrightarrow{d} N(0, V(\tau') - V(\tau^*)).$$

- Let $\hat{V}(\tau^*)$ be an estimate of $V(\tau^*)$ and similarly for $V(\tau')$. The Durbin-Wu-Hausman test statistic is given by

$$DWH = n(\hat{\theta}(\tau^*) - \bar{\theta}(\tau^*))'[\hat{V}(\tau') - \hat{V}(\tau^*)]^-(\hat{\theta}(\tau^*) - \bar{\theta}(\tau^*))$$

  where A^- denotes the Moore-Penrose pseudoinverse of A. Under H_0,

$$DWH \xrightarrow{d} \chi^2_r$$

  where $r = \text{rank}(V(\tau') - V(\tau^*))$ (typically r corresponds to the number of "restrictions" involved in constraining the model so that τ^* is efficient).

- Example: for testing exogeneity in linear IV with homoskedastic errors, the J-test ends up taking this form. See pp. 2235-2236 in NM.

References

Berry, S., J. Levinsohn, and A. Pakes (1999): "Voluntary Export Restraints on Automobiles: Evaluating a Trade Policy," *The American Economic Review*, 89, 400–430.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. E. a. D. L. McFadden, Elsevier, vol. 4, 2111–2245.

Reynaert, M. and F. Verboven (2014): "Improving the performance of random coef-

ficients demand models: The role of optimal instruments," *Journal of Econometrics*, 179, 83–98.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*, Cambridge, Mass: The MIT Press, second edition edition ed.

Wright, J. H. (2003): "Detecting Lack of Identification in Gmm," *Econometric Theory*, 19, 322–330.