GMM Estimation and Testing

Whitney Newey

October 2007

Idea: Estimate parameters by setting sample moments to be close to population counterpart.

Definitions:

$$\begin{aligned} \beta \quad &: \quad p \times 1 \text{ parameter vector, with true value } \beta_0. \\ g_i(\beta) \quad &= \quad g(w_i, \beta) : m \times 1 \text{ vector of functions of } i^{th} \text{ data observation } w_i \text{ and paramet} \end{aligned}$$

Model (or moment restriction):

$$E[g_i(\beta_0)] = 0.$$

Definitions:

$$\begin{aligned} \hat{g}(\beta) \overset{def}{=} \sum_{i=1}^{n} g_i(\beta)/n \quad &: \quad \text{Sample averages.} \\ \hat{A} \quad &: \quad m \times m \text{ positive semi-definite matrix.} \end{aligned}$$

GMM ESTIMATOR:

$$\hat{\beta} = \arg\min_{\beta} \hat{g}(\beta)' \hat{A} \hat{g}(\beta).$$

GMM ESTIMATOR:

$$\hat{\beta} = \arg\min_{\beta} \hat{g}(\beta)'\hat{A}\hat{g}(\beta).$$

Interpretation:

Choosing $\hat{\beta}$ so sample moments are close to zero.

For $\|g\|_{\hat{A}} = \sqrt{g'\hat{A}g}$, same as minimizing $\|\hat{g}(\beta) - 0\|_{\hat{A}}$.

When $m = p$, the $\hat{\beta}$ with $\hat{g}(\hat{\beta}) = 0$ will be the GMM estimator for any $\hat{A}$

When $m > p$ then $\hat{A}$ matters.

Method of Moments is Special Case:

$$\begin{aligned} \text{Moments} &: \ E[y^j] = h_j(\beta_0), (1 \le j \le p), \\ \text{Specify moment functions} &: \ g_i(\beta) = (y_i - h_1(\beta), ..., y_i^p - h_p(\beta))', \\ \text{Estimator} &: \ \hat{g}(\hat{\beta}) = 0 \text{ same as } \overline{y^j} = h_j(\hat{\beta}),\ (1 \le j \le p). \end{aligned}$$

Two-stage Least Squares as GMM:

$$\text{Model}: y_i = X_i'\beta + \varepsilon_i, E[Z_i\varepsilon_i] = 0, (i = 1, ..., n).$$

$$\text{Specify moment functions}: g_i(\beta) = Z_i(y_i - X_i'\beta).$$

$$\begin{aligned} \text{Sample moments are} \quad &: \quad \hat{g}(\beta) = \sum_{i=1}^{n} Z_i(y_i - X_i'\beta)/n = Z'(y - X\beta)/n \\ Z \quad &= \quad [Z_1, ..., Z_n]', X = [X_1, ..., X_n]', y = (y_1, ..., y_n)'. \end{aligned}$$

$$\text{Specify distance (weighting) matrix}: \hat{A} = (Z'Z/n)^{-1}.$$

$$\begin{aligned} \text{GMM estimator is 2SLS} \quad &: \quad \hat{\beta} = \arg\min_{\beta}[(y - X\beta)'Z/n](Z'Z/n)^{-1}Z'(y - X\beta)/n \\ &= \quad \arg\min_{\beta}(y - X\beta)'Z(Z'Z)^{-1}Z'(y - X\beta). \end{aligned}$$

Intertemporal CAPM

Nonlinear in parameters.

c_i consumption at time i, R_i is asset return between i and $i+1$, α_0 is time discount factor, $u(c, \gamma_0)$ utility function;

I_i is variables observed at time i;

Agent maximizes

$$E[\sum_{j=0}^{\infty} \alpha_0^{-j} u(c_{i+j}, \gamma_0)]$$

subject to intertemporal budget constraint.

First-order conditions are

$$E[R_i \cdot \alpha_0 \cdot u_c(c_{i+1}, \gamma_0)/u_c(c_i, \gamma_0)|I_i] = 1.$$

Marginal rate of substitution between i and $i+1$ equals rate of return (in expected value).

Let Z_i denote a $m \times 1$ vector of variables observable at time i (like lagged consumption, lagged returns, and nonlinear functions of these).

Moment function is

$$g_i(\beta) = Z_i\{R_i \cdot \alpha \cdot u_c(c_{i+1}, \gamma_0)/u_c(c_i, \gamma_0) - 1\}.$$

Here GMM is nonlinear instrumental variables.

Empirical Example: Hansen and Singleton (1982, Econometrica).

Dynamic Panel Data

It is a simple model that is important starting point for microeconomic (e.g. firm investment) and macroeconomic (e.g. cross-country growth) applications is

$$E[y_{it}|y_{i,t-1},y_{i,t-2},...,y_{i0},\alpha_i]=\beta_0 y_{i,t-1}+\alpha_i,$$

α_i is unobserved individual effect.

Microeconomic application is firm investment (with additional covariates).

Macroeconomic is cross-country growth equations.

Let $\eta_{it}=y_{it}-E[y_{it}|y_{i,t-1},...,y_{i0},\alpha_i]$.

$$\begin{aligned} E[y_{i,t-j}\eta_{it}] &= 0,(1\leq j\leq t,t=1,...,T), \\ E[\alpha_i\eta_{it}] &= 0,(t=1,...,T). \end{aligned}$$

Δ denote first difference, i.e. $\Delta y_{it} = y_{it} - y_{i,t-1}$, so $\Delta y_{it} = \beta_0 \Delta y_{i,t-1} + \Delta \eta_{it}$. Then

$$E[y_{i,t-j}(\Delta y_{it} - \beta_0 \Delta y_{i,t-1})] = 0, (2 \leq j \leq t, t = 1, ..., T).$$

IV moment conditions. Twice (and more) lagged levels of y_{it} can be used as instruments for differenced equations.

Different instruments for different residuals.

Additional moment conditions from orthogonality of α_i and η_{it}. They are

$$E[(y_{iT} - \beta_0 y_{i,T-1})(\Delta y_{it} - \beta_0 \Delta y_{i,t-1})] = 0, (t = 2, ..., T-1).$$

These are nonlinear.

Combine moment conditions by stacking. Let

$$\begin{aligned} g_i^t(\beta) &= \begin{pmatrix} y_{i0} \\ \vdots \\ y_{i,t-2} \end{pmatrix} (\Delta y_{it} - \beta \Delta y_{i,t-1}), (t = 2, ..., T), \\ g_i^\alpha(\beta) &= \begin{pmatrix} \Delta y_{i2} - \beta \Delta y_{i1} \\ \vdots \\ \Delta y_{i,T-1} - \beta \Delta y_{i,T-2} \end{pmatrix} (y_{iT} - \beta y_{i,T-1}). \end{aligned}$$

These moment functions can be combined as

$$g_i(\beta) = (g_i^2(\beta)', ..., g_i^T(\beta)', g_i^\alpha(\beta)')'.$$

Here there are $T(T-1)/2 + (T-2)$ moment restrictions.

Ahn and Schmidt (1995, Journal of Econometrics) show that the addition of the nonlinear moment condition $g_i^\alpha(\beta)$ to the IV ones often gives substantial asymptotic efficiency improvements.

Arellano and Bond approach also better is small samples:

$$g_i^t(\beta) = \begin{pmatrix} \Delta y_{i,1} \\ \Delta y_{i,2} \\ \vdots \\ \Delta y_{i,t-1} \end{pmatrix} (y_{it} - \beta y_{i,t-1})$$

Assumes that have representation

$$y_{it} = \sum_{j=1}^{\infty} a_{tj} y_{i,t-j} + \sum_{j=1}^{\infty} b_{tj} \eta_{i,t-j} + C\alpha_i.$$

Hahn, Hausman, Kuersteiner approach: Long differences

$$g_i(\beta) = \begin{pmatrix} y_{i0} \\ y_{i2} - \beta y_{i1} \\ \vdots \\ y_{i,T-1} - \beta y_{i,T-2} \end{pmatrix} (y_{iT} - y_{i1} - \beta(y_{i,T-1} - y_{i0}))]$$

Has better small sample properties by getting most of the information with fewer moment conditions.

IDENTIFICATION:

Identification precedes estimation.

If a parameter is not identified then no consistent estimator exists.

Identification from moment functions: Let $\bar{g}(\beta) = E[g_i(\beta)]$.

Identification condition is

$$\beta_0 \textbf{ IS THE ONLY SOLUTION TO } \bar{g}(\beta) = 0$$

Necessary condition for identification is that

$$m \geq p.$$

When $m < p$, i.e. there are fewer equations to solve than parameters, there will typically be multiple solutions to the moment conditions.

In IV need at least as many instrumental variables as right-hand side variables.

Rank condition for identification:

Let $G = E[\partial g_i(\beta_0)/\partial\beta]$. Rank condition is

$$rank(G) = p.$$

Necessary and sufficient for identification when $g_i(\beta)$ is linear in β.

$$0 = \bar{g}(\beta) = \bar{g}(\beta_0) + G(\beta - \beta_0) = G(\beta - \beta_0)$$

has a unique solution at β_0 if and only if the rank of G equals the number of columns of G.

For IV $G = -E[Z_i X_i']$ so that $rank(G) = p$ is usual rank condition.

In the general nonlinear case it is difficult to specify conditions for uniqueness of the solution to $\bar{g}(\beta) = 0$.

$rank(G) = p$ will be sufficient for local identification.

There exists a neighborhood of β_0 such that β_0 is the unique solution to $\bar{g}(\beta)$ for all β in that neighborhood.

Exact identification is $m = p$.

For IV as many instruments as right-hand side variables.

Here the GMM estimator will satisfy $\hat{g}(\hat{\beta}) = 0$ asymptotically; see notes.

Overidentification is $m > p$.

TWO STEP OPTIMAL GMM ESTIMATOR:

When $m > p$ GMM estimator depends on weighting matrix $\hat{A}$.

Optimal $\hat{A}$ minimizes asymptotic variance of $\hat{\beta}$.

Let Ω be asymptotic variance of $\sqrt{n}\hat{g}(\beta_0) = \sum_{i=1}^{n} g_i(\beta_0)/\sqrt{n}$, i.e.

$$\sqrt{n}\hat{g}(\beta_0) \xrightarrow{d} N(0, \Omega).$$

In general, central limit theorem for time series gives

$$\Omega = \lim_{n \longrightarrow \infty} E[n\hat{g}(\beta_0)\hat{g}(\beta_0)'].$$

Optimal $\hat{A}$ satisfies

$$\hat{A} \xrightarrow{p} \Omega^{-1}.$$

Take

$$\hat{A} = \hat{\Omega}^{-1}$$

for $\hat{\Omega}$ a consistent estimator of Ω.

Estimating Ω.

Let $\tilde{\beta}$ be preliminary GMM estimator (known $\hat{A}$).

Let $\tilde{g}_i = g_i(\tilde{\beta}) - \hat{g}(\tilde{\beta})$

If $E[g_i(\beta_0)g_{i+\ell}(\beta_0)'] = 0$ then for some pre

$$\hat{\Omega} = \sum_{i=1}^{n} \tilde{g}_i \tilde{g}_i{}'/n.$$

Example is CAPM model above.

If have autocorrelation in moment conditions then

$$\begin{aligned}\hat{\Omega} &= \hat{\Lambda}_0 + \sum_{\ell=1}^{L} w_{\ell L}(\hat{\Lambda}_\ell + \hat{\Lambda}'_\ell), \\ \hat{\Lambda}_\ell &= \sum_{i=1}^{n-\ell} \tilde{g}_i \tilde{g}'_{i+\ell}/n.\end{aligned}$$

Weights $w_{\ell L}$ make $\hat{\Omega}$ positive semi-definite.

Bartlett weights $w_{\ell L} = 1 - \ell/(L+1)$, as in Newey and West (1987).

Choice of L.

Two step optimal GMM:

$$\hat{\beta} = \arg\min_{\beta} \hat{g}(\beta)'\hat{\Omega}^{-1}\hat{g}(\beta).$$

Two step optimal instrumental variables: For $\hat{g}(\beta) = Z'(y - X\beta)/n$,

$$\hat{\beta} = (X'Z\hat{\Omega}^{-1}Z'X)^{-1}X'Z\hat{\Omega}^{-1}Z'y.$$

Two-stage least squares versus optimal GMM: Using $\hat{\Omega}^{-1}$ improves asymptotic efficiency but may be worse in small samples due to higher variability of $\hat{\Omega}$, that estimates fourth moments.

CONSISTENT ASYMPTOTIC VARIANCE ESTIMATION

Optimal two step GMM estimator has

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), V = (G'\Omega^{-1}G).$$

To form asymptotic t-statistics and confidence intervals need a consistent estimator $\hat{V}$ of V.

Let $\hat{G} = \partial\hat{g}(\hat{\beta})/\partial\beta$. A consistent estimator of V is

$$\hat{V} = (\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}.$$

Could also update $\hat{\Omega}$.

ASYMPTOTIC THEORY FOR GMM:

Consistency for i.i.d. data.

If the data are i.i.d. and i) $E[g_i(\beta)] = 0$ *if and only if* $\beta = \beta_0$ *(identification); ii) the GMM minimization takes place over a compact set* B *containing* β_0*; iii)* $g_i(\beta)$ *is continuous at each* β *with probability one and* $E[\sup_{\beta \in \mathcal{B}} \|g_i(\beta)\|]$ *is finite; iv)* $\hat{A} \xrightarrow{p} A$ *positive definite; then* $\hat{\beta} \xrightarrow{p} \beta_0$.

Proof in Newey and McFadden (1994).

Idea: $\hat{g}(\beta_0) \xrightarrow{p} 0$ by law of large numbers.

Therefore $\hat{g}(\beta_0)'\hat{A}g(\beta_0) \xrightarrow{p} 0$.

Also $\hat{g}(\hat{\beta})'\hat{A}\hat{g}(\hat{\beta}) \leq \hat{g}(\beta_0)'\hat{A}g(\beta_0)$, so

$$\hat{g}(\hat{\beta})'\hat{A}g(\hat{\beta}) \xrightarrow{p} 0.$$

The only way this can happen is if $\hat{\beta} \xrightarrow{p} \beta_0$.

Asymptotic normality:

If the data are i.i.d., $\hat{\beta} \xrightarrow{p} \beta_0$ and i) β_0 is in the interior of the parameter set over which minimization occurs; ii) $g_i(\beta)$ is continuously differentiable on a neighborhood N of β_0 iii) $E[\sup_{\beta\in\mathcal{N}} \|\partial g_i(\beta)/\partial\beta\|]$ is finite; iv) $\hat{A} \xrightarrow{p} A$ and $G'AG$ is nonsingular, for $G = E[\partial g_i(\beta_0)/\partial\beta]$; v) $\Omega = E[g_i(\beta_0)g_i(\beta_0)']$ exists, then

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), V = (G'AG)^{-1}G'A\Omega AG(G'AG)^{-1}.$$

See Newey and McFadden (1994) for the proof.

Can give derivation:

Let $\hat{G} = \partial \hat{g}(\hat{\beta})/\partial \beta$.

First order conditions are

$$0 = \hat{G}' \hat{A} \hat{g}(\hat{\beta})$$

Expand $\hat{g}(\hat{\beta})$ around β_0 to obtain

$$\hat{g}(\hat{\beta}) = \hat{g}(\beta_0) + \bar{G}(\hat{\beta} - \beta_0),$$

where $\bar{G} = \partial \hat{g}(\bar{\beta})/\partial \beta$ and $\bar{\beta}$ lies on the line joining $\hat{\beta}$ and β_0, and actually differs from row to row of $\bar{G}$.

Substitute this back in first order conditions to get

$$0 = \hat{G}' \hat{A} \hat{g}(\beta_0) + \hat{G}' \hat{A} \bar{G}(\hat{\beta} - \beta_0),$$

Solve for $\hat{\beta} - \beta_0$ and mulitply through by $\sqrt{n}$ to get

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left(\hat{G}' \hat{A} \bar{G}\right)^{-1} \hat{G}' \hat{A} \sqrt{n} \hat{g}(\beta_0).$$

Recall

$$\sqrt{n}(\hat{\beta} - \beta_0) = -\left(\hat{G}'\hat{A}\bar{G}\right)^{-1}\hat{G}'\hat{A}\sqrt{n}\hat{g}(\beta_0).$$

By central limit theorem

$$\sqrt{n}\hat{g}(\beta_0) \xrightarrow{d} N(0, \Omega)$$

Also we have $\hat{A} \xrightarrow{p} A$, $\hat{G} \xrightarrow{p} G$, $\bar{G} \xrightarrow{p} G$, so by the continuous mapping theorem,

$$\left(\hat{G}'\hat{A}\bar{G}\right)^{-1}\hat{G}'\hat{A} \xrightarrow{p} \left(G'AG\right)^{-1}G'A$$

Then by the Slutzky lemma.

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} -\left(G'AG\right)^{-1}G'AN(0, \Omega) = N(0, V).$$

The fact that $A = \Omega^{-1}$ minimizes the asymptotic varince follows from the Gauss Markov Theorem.

Consider a linear model.

$$E[Y] = G\delta, Var(Y) = \Omega.$$

$(G'\Omega^{-1}G)^{-1}$ is the variance of generalized least squares (GLS) in this model.

$V = (G'AG)^{-1}G'A\Omega AG(G'AG)^{-1}$ is the variance of $\hat{\delta} = (G'AG)^{-1}G'AY$.

GLS has smallest variance by Gauss-Markov theorem,

$$V - (G'\Omega^{-1}G)^{-1} \text{ is p.s.d.}$$

TESTING IN GMM:

Overidentification test statistic is

$$n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}).$$

Limiting distribution is chi-squared with $m-p$ degrees of freedom.

Test only of overidentifying restrictions.

What would value be if $m=p$.

"Use up" p restrictions in estimating β.

Will have no power against some misspecification which biases $\hat{\beta}$.

Testing subsets of moment conditions. Partition $g_i(\beta) = (g_i^1(\beta)', g_i^2(\beta)')'$ and Ω conformably. $H_0 : E[g_i^1(\beta)] = 0$. One simple test is

$$\hat{T}_1 = \min_\beta n\hat{g}(\beta)'\hat{\Omega}^{-1}\hat{g}(\beta) - \min_\beta n\hat{g}^2(\beta)'\hat{\Omega}_{22}^{-1}\hat{g}^2(\beta).$$

Asymptotic distribution of this is $\chi^2(m_1)$ where m_1 is the dimension of $\hat{g}^1(\beta)$.

Another version is

$$n\tilde{g}^{1\prime}\tilde{\Sigma}^{-1}\tilde{g}^1,$$

where $\tilde{g}^1 = \hat{g}^1(\hat{\beta}) - \hat{\Omega}_{12}\hat{\Omega}_{22}^{-1}\hat{g}^2(\hat{\beta})$ and $\tilde{\Sigma}^{-1}$ is estimator of asymptotic variance of $\sqrt{n}\tilde{g}^1$.

Hausman test: If $m_1 \leq p$ then, except in degenerate cases, Hausman test based on the difference of the optimal two-step GMM estimator using all the moment conditions and using just $\hat{g}^2(\beta)$ will be asymptotically equivalent to this test, for any m_1 parameters.

See Newey (1985, GMM Specification Testing, Journal of Econometrics.)

Tests of $H_0 : s(\beta) = 0$.

Let $\tilde{\beta} = \arg\min_{s(\beta)=0} n\hat{g}(\beta)'\hat{\Omega}^{-1}\hat{g}(\beta)$ be restricted estimator.

Wald test statistic:

$$W = ns(\hat{\beta})'[\partial s(\hat{\beta})/\partial\beta(\hat{G}'\hat{\Omega}^{-1}\hat{G})^{-1}\partial s(\hat{\beta})/\partial\beta]^{-1}s(\hat{\beta}).$$

Likelihood ratio (sum of squared residuals test statistic).

$$LR = n\hat{g}(\tilde{\beta})'\hat{\Omega}^{-1}\hat{g}(\tilde{\beta}) - n\hat{g}(\hat{\beta})'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}).$$

Lagrange mulitplier test statistic: For $\tilde{G} = \partial\hat{g}(\tilde{\beta})/\partial\beta$,

$$LM = n\hat{g}(\tilde{\beta})'\hat{\Omega}^{-1}\tilde{G}(\tilde{G}'\hat{\Omega}^{-1}\tilde{G})^{-1}\tilde{G}'\hat{\Omega}^{-1}\hat{g}(\tilde{\beta}).$$

ADDING MOMENT CONDITIONS:

Improves asymptotic efficiency but can lead to bias in two step GMM estimator. More on bias next time.

Efficiency improvment occurs because optimal weighting matrix for fewer moment conditions is not optimal for all the moment conditions.

Suppose that $g_i(\beta) = (g_i^1(\beta)', g_i^2(\beta)')'$. Then the optimal GMM estimator for just the first set of moment conditions $g_i^1(\beta)$ corresponds to a weighting matrix for all the moment conditions of the form

$$\hat{A} = \begin{pmatrix} (\hat{\Omega}_{11})^{-1} & 0 \\ 0 & 0 \end{pmatrix},$$

Not generally optimal for the entire moment function vector $g_i(\beta)$, where $\hat{\Omega}^{-1}$ is optimal.

Example: Linear regression model

$$E[y_i|X_i] = X_i'\beta_0.$$

Least squares estimator is GMM with $g_i^1(\beta) = X_i(y_i - X_i'\beta)$. Let $\rho_i(\beta) = y_i - X_i'\beta_0$.

More efficient estimator obtained by adding $g_i^2(\beta) = a(X_i)(y_i - X_i'\beta)$ for some $(m-p) \times 1$ vector of functions $a(X)$. GMM for

$$g_i(\beta) = \begin{pmatrix} X_i \\ a(X_i) \end{pmatrix} (y_i - X_i'\beta)$$

is more efficient than least squares with heteroskedasticity.

Example: Randomly missing data. Linear regression again. Some variables missing at random. W_i always observed.

Let $\Delta_i = 1$ denote a complete data indicator, equal to 1 if all variables observed, equal to zero otherwise. Complete data moment functions:

$$g_i^1(\beta) = \Delta_i X_i(y_i - X_i'\beta).$$

Add

$$g_i^2(\eta) = (\Delta_i - \eta)a(W_i).$$

No reduction in asymptotic variance when additional moment conditions are exactly identified.

Adding moment conditions lowers asymptotic variance but:

1) Increases small sample bias (with endogeneity present).

2) Increases small sample variance.

CONDITIONAL MOMENT RESTRICTIONS:

$\rho_i(\beta) = \rho(w_i, \beta)$ a $r \times 1$ residual vector. Instruments z_i such that

$$E[\rho_i(\beta_0)|z_i] = 0$$

Let $F(z_i)$ be an $m \times r$ matrix of functions of z_i (instrumental variables). Let $g_i(\beta) = F(z_i)\rho_i(\beta)$.

$$E[g_i(\beta_0)] = E[F(z_i)E[\rho_i(\beta_0)|z_{\beta i}]] = 0.$$

Can form GMM estimator using $g_i(\beta) = F(z_i)\rho_i(\beta)$. Leads to nonlinear IV estimator. Sargan (1958, 1959).

Optimal choice of $F(z)$: Let

$$D(z) = E[\partial\rho_i(\beta_0)/\partial\beta|z_i = z], \Sigma(z) = E[\rho_i(\beta_0)\rho_i(\beta_0)'|z_i = z].$$

Asymptotic variance minimizing $F(z)$ is

$$F^*(z) = D(z)'\Sigma(z)^{-1}.$$

Minimizes variance over all $F(z_i)$ a weighting matrices A. Note $F^*(z)$ is $p \times r$, so $g_i(\beta) = F^*(z_i)\rho_i(\beta)$ is $p \times 1$ (exact identification).

Examples:

Heteroskedastic Linear Regression: $E[y_i|X_i] = X_i'\beta_0$ and let $\rho_i(\beta) = y_i - X_i'\beta$.

GMM estimator has $g_i(\beta) = F(X_i)(y_i - X_i'\beta)$.

Here $z_i = X$ and $\partial\rho_i(\beta)/\partial\beta = -X_i$, so that

$$D(z_i) = -E[X_i|z_i] = -X_i, \mathbf{\Sigma}(z_i) = E[\rho_i(\beta_0)^2|X_i] = Var(y_i|X_i) = \sigma_i^2.$$

Optimal instruments:

$$F^*(z_i) = \frac{-X_i}{\sigma_i^2}.$$

GMM estimator solves

$$0 = \sum_i \frac{X_i}{\sigma_i^2}(y_i - X_i'\beta).$$

This is __________________.

Homoskedastic Linear Simultaneous Equation:

Here again $\rho_i(\beta) = y_i - X_i'\beta$ but now z_i is not X_i; z_i are instruments.

Assume $E[\rho_i(\beta_0)^2|z_i] = \sigma^2$ is constant.

Here $D(z_i) = -E[X_i|z_i]$ is the reduced form.

The optimal instruments in this example are

$$F = \frac{-D(z_i)}{\sigma^2} = -\sigma^{-2}E[X_i|z_i].$$

Here the reduced form may be nonlinear.

In general optimal instruments combine heteroskedasticity weighting with nonlinear reduced form.

Optimality proof:

Let $F_i = F(z_i)$ be instruments for some GMM estimator with moment function $g_i(\beta) = F(z_i)\rho_i(\beta)$, $F_i^* = F^*(z_i)$ be optimal instruments, and $\rho_i = \rho_i(\beta_0)$.

Iterated expectations gives

$$G = E[F_i \partial\rho_i(\beta_0)/\partial\beta] = E[F_i D(z_i)] = E[F_i \Sigma(z_i) F_i^{*\prime}] = E[F_i \rho_i \rho_i' F_i^{*\prime}].$$

For any weighting matrix A define $h_i = G'AF_i\rho_i$. Let $h_i^* = F_i^*\rho_i$. Then

$$G'AG = G'AE[F_i\rho_i h_i^{*\prime}] = E[h_i h_i^{*\prime}], G'A\Omega AG = E[h_i h_i'].$$

For $F_i = F_i^*$ we have $G = \Omega = E[h_i^* h_i^{*\prime}]$, so asymptotic variance of estimator with $F(z) = F^*$ is $(E[h_i^* h_i^{*\prime}])^{-1}$.

Variance of GMM estimator with F and A minus variance of estimator with F^* is

$$\begin{aligned} & (G'AG)^{-1}G'A\Omega AG(G'AG)^{-1} - \left(E[h_i^* h_i^{*\prime}]\right)^{-1} \\ = & \left(E[h_i h_i^{*\prime}]\right)^{-1} \{E[h_i h_i'] - E[h_i h_i^{*\prime}] \left(E[h_i^* h_i^{*\prime}]\right)^{-1} E[h_i^* h_i']\} \left(E[h_i^* h_i']\right)^{-1}. \end{aligned}$$

Approximating the optimal moment functions

For given $F(z)$ the GMM estimator with optimal weight matrix $A = \Omega^{-1}$ is approximation to the optimal estimator. For simplicity consider one residual, one parameter case, $r = p = 1$.

Let $g_i = F_i\rho_i$, as above, $G = E[g_i h_i^{*\prime}]$, so that

$$G'\Omega^{-1} = E[h_i^* g_i'](E[g_i g_i'])^{-1}.$$

$G'\Omega^{-1}$ are the coefficients of the population regression of h_i^* on g_i.

First-order conditions for GMM

$$0 = \hat{G}'\hat{\Omega}^{-1}\hat{g}(\hat{\beta}) = \hat{G}'\hat{\Omega}^{-1}\sum_{i=1}^{n} F(z_i)\rho_i(\hat{\beta})/n,$$

Lowest mean-square error approximation to

$$0 = \sum_{i=1}^{n} F_i^*\rho_i(\beta)/n.$$

Thus, if $F(z) = (a_{1m}(z), ..., a_{mm}(z))'$ such that linear combinations

$$\min_{\pi_1,\ldots,\pi_m} E[\{h_i^* - \pi' g_i\}^2] = \min_{\pi_1,\ldots,\pi_m} E[\Sigma(z_i)\{F_i^* - \pi' F(z_i)\}^2] \longrightarrow 0,$$

as $m \longrightarrow \infty$ then get variance of GMM approaches lower bound. An important issue for practice is the choice of m.