Clustering, random effects, mixed models, sandwiches

November 28, 2013

Grouped data

Notation: When a variable z_{ig} is indexed by i and g, $z_{\cdot g}$ refers to the vector consisting of all the z_{ig}'s with the given value of g.

$$y_{ig} = X_{ig}\beta + \varepsilon_{ig}\,, \quad i = 1, \ldots, n, \quad g = 1, \ldots, M\,.$$

g indexes groups (states, gender, age brackets, etc.), i indexes observations (people, years, firms, etc.).

GLS: two step

Assume residual correlation within groups, but not between:

$$\mathrm{Var}(\varepsilon_{\cdot g}) = \Omega, \text{ all } g\,, \qquad \mathrm{Cov}(\varepsilon_{\cdot g}, \varepsilon_{\cdot h}) = 0\,.$$

Estimate by OLS on all data, use the estimated β to form $\hat{\varepsilon} = y - X\beta$, estimate Ω as

$$\hat{\Omega} = \sum_{g=1}^{M} \hat{\varepsilon}_{\cdot g}\hat{\varepsilon}_{\cdot g}\,.$$

Then the full $nM \times nM$ $E[\varepsilon\varepsilon']$ matrix is block diagonal, with copies of Ω down the diagonal. The estimated Ω converges in probability to the true value as $g \to \infty$ by the law of large numbers, so we can be use it for feasible GLS estimation.

GLS: Likelihood approach

Assume normality, write down the likelihood for the sample, base inference on it:

$$|\Omega|^{-M/2}(2\pi)^{-Mn/2}e^{-\frac{1}{2}\sum_{g=1}^{M}(y_{\cdot g}-X_{\cdot g}\beta)'\Omega^{-1}(y_{\cdot g}-X_{\cdot g}\beta)}.$$

Because Ω is symmetric, it contains $(n^2+n)/2$ coefficients.

Clustered covariance matrix for β.

Assume

$$E[X'\varepsilon\varepsilon'X] = E\big[\sum_g X'_{\cdot g}\varepsilon_{\cdot g}\varepsilon'_{\cdot g}X_{\cdot g}\big]\,.$$

That is, no correlation of $X'_{\cdot g}\varepsilon_{\cdot g}$ with $X'_{\cdot h}\varepsilon_{\cdot h}$ for $g \neq h$, but arbitrary covariance, constant across g, when $g = h$.

Then apply the usual "robust" standard error form:

$$\mathrm{Var}(\hat{\beta}_{OLS}) \doteq (X'X)^{-1}E[X'\varepsilon\varepsilon'X](X'X)^{-1}\,,$$

Replacing the central expectation with

$$\sum_{g=1}^{M} X'_{\cdot g}\hat{\varepsilon}_{\cdot g}\hat{\varepsilon}'_{\cdot g}X_{\cdot g}.$$

This is called a **clustered** robust covariance matrix.

Which: OLS, GLS 2-step, GLS Likelihood, OLS with sandwich?

- A very common set of trade-offs here.
- Straight OLS with $\sigma^2(X'X)^{-1}$ covariance matrix is the most accurate and efficient if its assumptions are correct.
- Of the estimates that allow for $\mathrm{Var}(\varepsilon_{\cdot g}) \neq \sigma^2 I$, OLS with sandwich is easiest on the researcher's brain. No need to think about a structure for Ω or to defend the assumption of $E[\varepsilon_{\cdot g}\varepsilon_{\cdot g}]$ constant across groups.
- OLS with sandwich is a little more algebra than straight OLS with $\sigma^2(X'X)^{-1}$ covariance matrix, but the computer does that with a single button.

- Drawback: If there is a non-scalar covariance matrix for ε, one can do a better job of estimation, obtaining more precise results, by modeling the form of the covariance matrix.

- The sandwich estimator replaces clear assumptions that justify the procedure with untestable claims that approximations that work well when $g \to \infty$ are reliable in the current sample. (This is true of any appeal to asymptotic theory.)

Which: OLS, GLS 2-step, GLS Likelihood, OLS with sandwich?

- Likelihood-based GLS, if its assumptions are correct and the residual covariance matrix is non-scalar, is more efficient than OLS and also provides a correct distribution for β in finite samples.

Which: OLS, GLS 2-step, GLS Likelihood, OLS with sandwich?

- Likelihood-based GLS, if its assumptions are correct and the residual covariance matrix is non-scalar, is more efficient than OLS and also provides a correct distribution for β in finite samples.

- Of course, as with straight OLS, its assumptions need not be correct.

- Two-step GLS is a little easier computationally, and has the same asymptotic distribution as likelihood-based GLS. It does not have the same finite-sample justification. It is likely to provide a good starting point for iterative estimation and MCMC study of the likelihood in the likelihood-based framework.

GLS with sandwich?

GLS here assumes $E[\varepsilon_{\cdot g}\varepsilon'_{\cdot g}] \mid X \equiv \Omega$. In other words, that the covariance matrix of residuals within a group does not depend on X. So a sandwich covariance matrix for the GLS estimate of β might differ from what GLS delivers, even asymptotically. The usual tradeoffs are here — robustness against deviation from the GLS assumptions, vs. robustness against inapplicability of asymptotic theory in this sample.

Group-specific shifts

On this topic, these notes are closer to the way Lancaster sets out this topic than to the way I did it in class.

Same grouped data model, with one change:

$$y_{ig} = X_{ig}\beta + \nu_g + \varepsilon_{ig}\,, \quad i = 1, \ldots, n, \quad g = 1, \ldots, M\,.$$

What's new is the ν_g, a "disturbance" that changes all observations within a group by the same amount. I've used a greek letter for it, which makes it seem natural to treat it as part of the error term.

Applying GLS

If we assert that

$$\varepsilon \mid X \sim N(0, \sigma^2 \underset{Mn \times Mn}{I}), \qquad \nu \mid X \sim N(0, \tau^2 \underset{M \times M}{I}) \,,$$

and ε, ν independent, then this is a special case of grouped data with non-scalar covariance matrix. Instead of an unconstrained residual covariance matrix Ω for data within each group, we have the parametric form

$$\Omega = \sigma^2 I + \tau^2 \underset{n \times n}{\mathbf{1}} \,.$$

We could stop here, simply referring back to the discussion of likelihood-based GLS, but it is worth noting that there is an analogue of weighted least squares available because of the special structure of Ω.

Between and within regressions

As usual, if we can find a matrix W such that $W'\Omega W = I$, then OLS on the transformed data $X^*_{\cdot g} = W'X_{\cdot g}$, $y^*_{\cdot g} = W'y_{\cdot g}$ is equivalent to GLS. With this group-effect Ω matrix, we can choose W to have the symmetric form $W = \sigma^{-1}(I - \frac{1}{n}\mathbf{1}) + \delta\mathbf{1}$. It is possible to compute δ from σ^2 and τ^2, but this involves messy algebra. What matters for our purposes is that there is a symmetric inverse square root W of Ω of this form, and that

$$\delta \xrightarrow[\tau^2 \to 0]{} \frac{1}{\sigma n}$$

$$\delta \xrightarrow[\tau^2 \to \infty]{} 0\,.$$

Between and within regressions

Note that with this choice of W, $X^* = W'X_{\cdot g} = \sigma^{-1}\tilde{X}_{\cdot g} + \delta\bar{X}_{\cdot g}$, where $\bar{X}$ is a matrix in which each row is the mean of $X_{\cdot g}$ within group g and $\tilde{X}_{\cdot g}$ is the deviation of $X_{\cdot g}$ from its group mean. Note also that $\tilde{X}'_{\cdot g}\bar{X}_{\cdot g} = 0$, because the columns of $\bar{X}_{\cdot g}$ are constant and the sum of each column of $\tilde{X}_{\cdot g}$ is zero.

Therefore

$$X^{*\prime}X^* = \sigma^{-2}\tilde{X}'\tilde{X} + \delta^2\bar{X}'\bar{X}\,, \qquad X^{*\prime}y^* = \tilde{X}'\tilde{y} + \bar{X}'\bar{y}\,,$$

where the $*$'d vectors and matrices are of full length Mn, consisting of the grouped data stacked up vertically.

Between and within regressions

This lets us write the GLS estimator as a matrix weighted average of the OLS estimator $\hat{\beta}_w$ using $\tilde{X}, \tilde{y}$, called the "**within**" regression, and the OLS estimator $\hat{\beta}_b$ using the group means, called the "**between**" regression:

$$\hat{\beta}_{GLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^* = (\sigma^{-2}\tilde{X}'\tilde{X}+\delta^2\bar{X}'\bar{X})^{-1}(\sigma^{-2}\tilde{X}'\tilde{X}\hat{\beta}_w+\delta^2\bar{X}'\bar{X}\hat{\beta}_b).$$

This decomposition is of some use to programmers and to you if you try to look at small data sets with a calculator. But the important insight from it is that the GLS estimator, which is the classic **random effects** estimator, becomes the ordinary OLS estimator in the limit as τ^2 becomes very small (as we might have expected) and becomes purely the "within" regression in the limit as τ^2 gets very big. But this pure within regression is also what is known as the **fixed effects** estimator.

Fixed effects

The fixed effects estimator is what emerges if we assert dogmatically that the variance of the group means $\nu_{\cdot g}$ is infinite — i.e. we put a flat prior on ν_g. It is also, as we verified above, the result of using data on deviations from group means in an OLS estimation. And finally, it is also the result if we estimate by OLS the equation

$$y_{ig} = c_g + X_{ig}\beta + \varepsilon_{ig} \ ,$$

treating the c_g's (a new name for ν_g) as parameters to be estimated along with β.

Fixed vs. random effects

- Fixed effects always estimates a more dispersed (higher variance) distribution of c_g's across groups than the true distribution of c_g's, and this does not get better as the number of groups increases.

- Fixed effects requires giving up any attempt to estimate coefficients of variables that are constant within groups. Random effects models can do so, because they exploit the assumption that ν_g and X are uncorrelated.

- Fixed effects gives consistent estimates of β as $M \to \infty$, even if ν_g and $X_{.g}$ are correlated, while random effects does not.

Random effects correlated with X

$$y_{ig} = X_{ig}\beta + c_g + \varepsilon_{ig}\,, \qquad E[\varepsilon_{\cdot g} \mid X_{\cdot g}] = 0 \tag{1}$$

$$X_{\cdot g} = \gamma c_g + \tilde{X}_{\cdot g}\,, \qquad E[\tilde{X}_{\cdot g} \mid c_g] = 0 \qquad \text{or} \tag{2}$$

$$c_g = \vec{X}_{\cdot g}\psi + \xi_g\,, \qquad E[\xi_g \mid X_{\cdot g}] = 0\,. \tag{3}$$

Wooldridge calls this the "Chamberlain-Mundlak device". Using (3), we can substitute into (1) to obtain

$$y_{ig} = X_{ig}\beta + \vec{X}_{\cdot g}\psi + \xi_g + \varepsilon_{ig}$$

Random effects correlated with X

$$y_{ig} = X_{ig}\beta + \vec{X}_{.g}\psi + \xi_g + \varepsilon_{ig}$$

In this equation, $\vec{X}_{.g}$ is a single vector of length nK containing all the values of X_{ig} that occur in the group. It is the same vector for every i in the group. This is nk additional parameters. That is still usually smaller than the number of c_g parameters that enter the straight fixed-effects estimator. In some applications it might be reasonable to claim that the correlation of ν_g with the $\vec{X}_{.g}$ vector should be only via the group means of the X's. In that case the $\vec{X}_{.g}$ vector could be replaced by the $\bar{X}_{.g}$ vector, making the number of extra parameters much smaller.

Random effects correlated with X

$$y_{ig} = X_{ig}\beta + \vec{X}_{.g}\psi + \xi_g + \varepsilon_{ig}$$

This is an equation that can be estimated by standard grouped-data GLS. It does allow consistent estimation of $\mathrm{Var}(\nu_g)$, but it does not allow estimation of coefficients of possible $X_{.g}$ variables that are constant within groups, because for such variables the corresponding columns of $X_{.g}$ and $\vec{X}_{.g}$ are identical, and thus collinear.

Mixed models

This term does not have a precise and widely accepted definition, but generally refers to models in which not only the constant, but also coefficients of $X_{.g}$ variables, are allowed to be random and vary with g. The general form, assuming the constant vector is treated as part of the $X_{.g}$ matrix, is

$$y_{ig} = X_{ig}\beta_g + Z_{ig}\gamma + \varepsilon_{ig}$$
$$E[\varepsilon \mid X, Z] = 0$$
$$E[\beta_g \mid X, Z] = \bar{\beta}]$$

Mixed models

$$y_{ig} = X_{ig}\beta_g + Z_{ig}\gamma + \varepsilon_{ig}$$
$$E[\varepsilon \mid X, Z] = 0$$
$$E[\beta_g] \mid X, Z] = \bar{\beta}]$$

Some assumption on the joint distribution of β_g and ε_{ig} is needed. A common choice would be to make β_g and ε_{ig} jointly normal and independent of each other, with the full $Mn \times 1$ ε_{ig} vector $N(0, \sigma^2 I)$ and

$$\mathrm{Var}(\beta_g) = I \otimes \Sigma_\beta \,,$$

where Σ_β is an unknown and unrestricted covariance matrix and the "$A \otimes B$" notation refers to a Kronecker product.

Kronecker product

$A \otimes B$, where A is $m \times n$ and B is $p \times q$, is an $mp \times nq$ matrix consisting of mn $p \times q$ blocks, with the block in the i'th row position and j'th column position equal to $a_{ij}B$. For example

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \qquad B = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad \Rightarrow$$

$$A \otimes B = \begin{bmatrix} 2 & 4 \\ 1 & 2 \\ 6 & 8 \\ 3 & 4 \end{bmatrix}.$$

Why mixed models?

They give a fully articulated probability model for the data, and thus a likelihood function, while addressing the possibility that $E[\varepsilon_{\cdot g}\varepsilon'_{\cdot g} \mid X_{\cdot g}]$ might depend on $X_{\cdot g}$. This possibility is what clustered standard errors allow for that GLS does not. Clustered standard errors allow any kind of dependence between $X_{\cdot g}$ and $\varepsilon_{\cdot g}\varepsilon'_{\cdot g}$, while mixed models restrict the dependence to linearity.

Mixed models used to be intractably difficult to estimate, but with Gibbs sampling MCMC, they can be handled in a very straightforward way.

It might be a good way for you to test your understanding of Gibbs sampling to see if you can describe a convenient Gibbs sampling scheme for a mixed model.

"Fixed effects" for mixed models?

Mixed models almost always are used treating β_g as random. But there is an analogue to the simple fixed effects approach for these models: just as we split the constant into a bunch of group-dummy variables for fixed effects, we can split up the X matrix into a block diagonal form with $X_{\cdot g}$ blocks down the diagonal and zeros off diagonal, giving each column of this matrix its own free parameter and applying OLS.

The problem with this is that if there are very many columns in X, the fact that OLS with fixed effects allocates too much explanatory power to the fixed effects is multiplied in a mixed model — here it is not only the constant terms, but the coefficients on all the group-specific variables, that have an over-dispersed distribution.

GMM

Generalized Method of Moments

A large class of estimators

When the parameter β takes on its true value,

$$E[g(y_t, \beta)] = 0$$
$$\mathrm{Cov}\big(g(y_t, \beta)\big) = \Sigma_g \, .$$

The first of these, the moment condition, does not hold when β not at its true value.

A large class of estimators

Then why not estimate β from a sample $\{y_1, \ldots, y_T\}$ by solving

$$g_T(Y_T, \hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} g(y_t, \hat{\beta}) = 0 \, ?$$

In other words, solve to set the sample average of $g(y_t, \beta)$ equal to its theoretical population value.

A large class of estimators

Then why not estimate β from a sample $\{y_1, \ldots, y_T\}$ by solving

$$g_T(Y_T, \hat{\beta}) = \frac{1}{T} \sum_{t=1}^{T} g(y_t, \hat{\beta}) = 0 \, ?$$

In other words, solve to set the sample average of $g(y_t, \beta)$ equal to its theoretical population value.

This is GMM.

GMM covers many cases

- MLE:

$$g(y_t, \beta) = \frac{\partial \log p(y_t \mid \beta)}{\partial \beta}$$

- OLS:

$$g(y_t, \beta) = X_t'(y_t - X_t\beta)$$

- IV:

$$g(y_t, \beta) = Z_t'(y_t - X_t\beta)$$

What if the $g(y_t, \beta)$ vector is of length $m > k$, the length of the β vector?

Then, mechanically, we need a $k \times m$ weighting matrix W_T, which need not be the same in every sample size, so we can solve

$$W_T g_T(Y_T, \beta) = 0 \, .$$

Any weighting matrix gives us some version of GMM.

What is the best weighting matrix?

Asymptotic covariance matrix of GMM, i.i.d. case

Take Taylor expansion of the equation to be solved around the true value β_0 :

$$W_T g_T(Y_T, \beta) \doteq W_T g_T(y_T, \beta_0) + W_T \frac{\partial g_T(Y_T, \beta_0}{\partial \beta}(\beta - \beta_0) \, .$$

This gives us the approximate solution for $\beta - \beta_0$ as

$$\beta - \beta_0 = -(W_T \frac{\partial g_T(Y_T, \beta_0)}{\partial \beta})^{-1} W_T g_T(Y_T, \beta) \, .$$

Asymptotic covariance matrix

$$\frac{1}{T}\frac{\partial g_T(Y_T, \beta_0)}{\partial \beta} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial g(y_t, \beta)}{\partial \beta} \xrightarrow[t\to\infty]{P} A \,,$$

by the LLN, where A is some limiting matrix. We need to assume A is of full column rank. Also assume $W_T \xrightarrow[t\to\infty]{P} W$. The CLT tells us that

$$\frac{1}{\sqrt{T}} g_T(Y_T, \beta) \xrightarrow{D} N(0, B) \,, \qquad \text{where } B = E[g(y_t, \beta_0) g(y_t, \beta_0)'] \,.$$

Asymptotic covariance matrix

This lets us conclude that

$$\sqrt{T}(\hat{\beta}_T - \beta) \xrightarrow{D} N\big(0, (WA)^{-1}WBW'(A'W')^{-1}\big) .$$

But we get to choose W. How to do it so as to make this matrix "small"?

Optimal W

We'd like to make the covariance matrix "small" in the sense that with anything other than the optimal W, the covariance matrix differs from our optimal one by a positive semi-definite matrix. In other words, every linear combination of β's has either the same or lower asympotic variance. This is the same as making the inverse of the covariance matrix as big as possible, in the same sense of "bigger".

The inverse of the limiting covariance matrix is

$$A'W'(WBW')^{-1}WA\,.$$

Optimal W

The covariance matrix of the "explained sum of squares" in a regression of the columns of a matrix Y on X is $Y'X(X'X)^{-1}X'Y$. If we choose $X = B^{1/2}W'$ and $Y = B^{-1/2}A$, this explained some of squares expression matches our inverse covariance matrix. That is, the inverse of the covariance matrix is just the explained sum of squares from a regression of $B^{-1/2}A$ on $B^{1/2}W'$. ($B^{1/2}$ can be any square root of the B matrix. We're taking it to be the symmetric square root.)

Obviously the explained sum of squares is biggest when we make $X = Y$, which we can do here by choosing W to make

$$B^{1/2}W' = B^{-1/2}A\,, \qquad \therefore W' = B^{-1}A\,.$$

This makes the limiting covariance matrix $(A'B^{-1}A)^{-1}$.

GMM derived as a minimization

The GMM estimator can be defined as one that minimizes

$$g_T(\beta)'\Omega^{-1}g_T(\beta).$$

The first-order condition for a minimum is

$$\left(\frac{\partial g_T(Y_T,\hat{\beta})}{\partial \beta}\right)' \Omega^{-1} g_T(Y_T,\hat{\beta}) = 0\,.$$

If $g()$ (and hence Ω) are of the same dimension as β and $\partial g_T/\partial\beta$ and Ω are non-singular, this equation is solved if and only if $g_T(Y_T,\hat{\beta}) = 0$, so the estimator is just the GMM estimator as we have previously defined it.

When g is of higher dimension than β, the first-order condition defines a GMM estimator in our previous notation, with a weighting matrix

$$W_T = \left(\frac{\partial g_T(Y_T, \beta_T}{\partial \beta} \right)' \Omega^{-1} .$$

We know that the optimal limiting form for W_T is $W' = B^{-1}A$, while this W_T from the minimization converges to $A'\Omega^{-1}$. So to get optimal weighting we must set

$$\Omega = B = E[g(y_t, \beta)g(y_t, \beta)'] .$$

ML as GMM

Suppose y_t is i.i.d. with pdf $p(y_t \mid \theta)$. Then

$$E\left[\frac{\partial \log p(y_t \mid \theta)}{\partial \theta} \mid \theta\right] = 0\,.$$

This follows because, when we write the expression out as an integral, it becomes

$$\int \frac{1}{p(y_t \mid \theta)} \frac{\partial p(y_t \mid \theta)}{\partial \theta} p(y_t \mid \theta)\, dy_t = \frac{\partial}{\partial \theta} \int p(y_t \mid \theta)\, dy_t = 0\,.$$

The last equality follows because a pdf always integrates to one, and $p()$ is a pdf for y for every θ.

ML as GMM

Now we can set $g(y_t, \theta)$ in the GMM formulas equal to $\partial \log p(y_t \mid \theta)/\partial\theta$, and our previous derivations of the asymptotic normal distribution for the GMM estimator apply. Since there are exactly as many moment conditions as parameters (one partial derivative per parameter), we don't have to worry about optimal weighting. The usual $A^{-1}B(A^{-1})'$ expression for the asymptotic covariance matrix of GMM with no weighting applies, where here

$$A = E\left[\frac{\partial^2 \log p(y_t \mid \theta)}{\partial\theta\partial\theta'}\right]$$

$$B = E\left[\frac{\partial \log p(y_t \mid \theta)}{\partial\theta}\left(\frac{\partial \log p(y_t \mid \theta)}{\partial\theta}\right)'\right] .$$

ML as GMM

As a final simplification, observe that

$$E\left[\frac{\partial \log p(y_t \mid \theta)}{\partial \theta}\left(\frac{\partial \log p(y_t \mid \theta)}{\partial \theta}\right)'\right] = \int \frac{1}{p(y_t \mid \theta)} \frac{\partial^2 p(y_t \mid \theta)}{\partial \theta \partial \theta'} p(y_t \mid \theta)\, dy_t$$
$$- \int \left(\frac{1}{p(y_t \mid \theta)} \frac{\partial p(y_t \mid \theta)}{\partial \theta}\right)\left(\frac{1}{p(y_t \mid \theta)} \frac{\partial p(y_t \mid \theta)}{\partial \theta}\right)' p(y_t \mid \theta)\, dy_t = 0 + B\ .$$

That is, in the maximum likelihood case, so long as the likelihood is in fact the pdf of the observed data, $-B = A$. The fact that the first term in the expression above is zero follows because it is the second derivative with respect to θ of the integral of $p(y_t \mid \theta)$ with respect to y_t — which is one, for all θ.

ML as GMM

Thus we arrive at the conclusion that the asymptotic covariance matrix of the MLE is

$$\left(-E \left[\frac{\partial^2 \log p(y_t \mid \theta)}{\partial \theta \partial \theta'} \right] \right)^{-1}$$

This can be consistently estimated, in the i.i.d. case, as the sample average of these second derivative matrices.

ML plus sandwich?

The matrix we have been calling B can be estimated directly as the sample average of the crossproducts of the scores (the derivatives of the log likelihood w.r.t. the parameters). In the case of ML, as we have just observed in the previous slide, we can instead use the fact that B is also the expected value of minus the second derivative of the log likelihood. This expected second derivative matrix can be found analytically as a function of the parameters θ, which are generally much smaller in number than the elements of the matrix of cross-products of scores. Therefore it is more efficient, maybe much more effficient, to use the second derivative matrix, plugging in estimated values of θ, rather than the estimate based on averaging the sample scores.

ML plus sandwich?

However, this depends on trusting the model to be correct when θ is at its true value. If the likelihood function is wrong, the MLE $\hat{\theta}_T$ still converges to some well-defined probability limit θ_∞ in regular cases, and it can be shown that there is a sense in which $p(y_t \mid \theta_\infty)$ must be as close as possible to the true pdf — but it is not the true pdf for y_t. If it is not, then using the full sandwich $A^{-1}BA^{-1}$ gives the correct asymptotic covariance matrix, while the $(-A)^{-1}$ efficient estimator is incorrect.

This is the familiar tradeoff: when the efficient estimate based on trusting the model is different from the "asymptotically robust" estimator for the covariance matrix, do we use the robust estimator, or do we take the difference as a criticism of our model and modify the model? This is the same issue that arises in deciding whether to use an HCCM covariance matrix or $\sigma^2(X'X)^{-1}$ in OLS.

Thinking about instrumental variables

[For this topic, see the paper "Thinking About Instrumental Variables", on my web site.] The main theme of that paper is that, when using asymptotic distribution theory for inference, it is useful where possible to examine the likelihood function for a model that would exactly justify working with the statistics (i.e., functions of the data) that enter the asymptotic distribution theory. For linear IV or 2SLS models, these statistics are just first and second moments of the endogeneous variables Y and X and the instrumental variables Z. The distribution that makes these sufficient statistics (and hence implies that we lose nothing by basing our inference entirely on them) is a joint normal distribution. With the joint normality assumption, we can construct a likelihood function. Maximizing that likelihood produces what is known as the limited information maximum likelihood (**LIML**) estimator, which has the same asymptotic distribution as 2SLS, but is different in finite samples.

Thinking about instrumental variables

Using the likelihood function is helpful in handling the two "breakdown cases" for 2SLS: weak instruments and embarrassingly many instruments. With weak instruments, the likelihood becomes very non-Gaussian if there is much probability weight on low R^2 in the regression of X (right-hand-side variables) on Z (instruments). Whether this is a problem and invalidates usual asymptotic approximations can be checked in a particular sample by examining the likelihood function to see whether its non-Gaussian character shows up in regions of high posterior probability.

With embarrassingly many instruments, the regression of X on Z produces R^2's of one, or nearly one, and we know this is from there being few degrees of freedom, not because the true R^2 is that high. 2SLS collapses to OLS in this case, while likelihood-based inference continues to give useful results.

Thinking about GMM

The ideas in "Thinking about instrumental variables" apply in principle also to GMM, but with GMM there is no general, automatic way to arrive at a model whose sufficient statistics match those generating the GMM estimator. Usually, converting GMM estimates to statements about a probability distribution for the unknown parameters θ must rely on the large-sample approximation that $\sqrt{T}(\theta - \hat{\theta}) \sim N(0, \Sigma)$, and that this assertion holds true whether we treat it as an assertion about the distribution of $\hat{\theta} \mid \theta$ (frequentist) or about the distribution of $\theta \mid \hat{\theta}$ (Bayesian).