LECTURE NOTE 9

SYSTEM ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

1. INTRODUCTION

We shall work with the following general framework:

$$Q_i = R_i \gamma + V_i, \tag{1}$$

$$E(W_i V_i) = 0, \tag{2}$$

where we observe (Q_i, R_i, W_i) for $i = 1, \ldots, n$. As in Note 6, we shall assume random sampling, so that (Q_i, R_i, W_i) are independent and identically distributed (i.i.d.) from some joint distribution. The system aspect is that Q_i can be a vector: Q_i is $H \times 1$, R_i is $H \times K$, the parameter vector γ is $K \times 1$, and the vector of errors V_i is $H \times 1$. Estimation of γ is based on the orthogonality between W_i and the error V_i. The matrix W_i is $L \times H$, so $W_i V_i$ is a $L \times 1$ vector, and $E(W_i V_i) = 0$ provides L orthogonality conditions to estimate the K components of γ. We need $L \geq K$. The variables in the matrix W_i are sometimes called "instrumental variables." In the current context, this just means that they generate orthogonality conditions.

A single linear predictor is a special case of this framework. Let $H = 1$ (not a system) and set $Q_i = Y_i$, $R_i = X_i'$, $W_i = X_i$. Then

$$Y_i = X_i' \gamma + V_i, \qquad E(X_i V_i) = 0$$

is equivalent to

$$E^*(Y_i \,|\, X_i) = X_i' \gamma.$$

The results obtained in this note will include the limit distribution for the least-squares estimator (Note 8) as a special case. In working with a single linear predictor, we saw in Note 8 that we can do inference (in large samples) without making assumptions about the conditional variance of Y_i given X_i—we can allow for heteroskedasticity of a general, unknown form. That is still true in the system framework, but now there is an additional issue. The components of the error vector V_i may be correlated with each other, and our inference methods need to allow for that. The next section shows how our panel data models from Notes 4 and 5 can fit into the system framework. With longitudinal panel data, the correlation between components of V_i comes from correlation over time (serial correlation) in the errors for the same cross-section unit.

2. PANEL DATA

2.1 *Complete Conditioning*

First consider the case with complete conditioning, as in Note 4:

$$E(Y_{it} \mid Z_{i1}, \ldots, Z_{iT}, A_i) = Z_{it}'\gamma + A_i \qquad (i = 1, \ldots, N;\ t = 1, \ldots, T). \tag{3}$$

For example, Y_{it} is the log of output for firm i in year t and Z_{it} is a $K \times 1$ vector containing measured inputs, such as capital and labor, for firm i in year t. The variable A_i is not observed; it is a firm effect that captures unmeasured inputs and differences in productivity that are constant over time, with variation only across the firms. This model for the conditional expectation imposes the exclusion restriction that only Z_{it} matters once we control for A_i—the past and future values of the measured inputs are excluded.

Define a prediction error:

$$U_{it} = Y_{it} - E(Y_{it} \mid Z_{i1}, \ldots, Z_{iT}, A_i), \tag{4}$$

so we can write the equations

$$Y_{it} = Z_{it}'\gamma + A_i + U_{it} \qquad (t = 1, \ldots, T).$$

The unmeasured variable A_i can create omitted variable bias. We can eliminate A_i by taking deviations from the time averages for each cross-section unit:

$$\bar{Y}_i \equiv \frac{1}{T}\sum_{t=1} Y_{it} = \bar{Z}_i'\gamma + A_i + \bar{U}_i,$$

$$Y_{it} - \bar{Y}_i = (Z_{it} - \bar{Z}_i)'\gamma + (U_{it} - \bar{U}_i) \quad (t = 1, \ldots, T). \tag{5}$$

The T deviation equations in (5) provide a system that maps into our framework in (1). Let

$$Q_i = \begin{pmatrix} Y_{i1} - \bar{Y}_i \\ \vdots \\ Y_{iT} - \bar{Y}_i \end{pmatrix}, \quad R_i = \begin{pmatrix} (Z_{i1} - \bar{Z}_i)' \\ \vdots \\ (Z_{iT} - \bar{Z}_i)' \end{pmatrix}, \quad V_i = \begin{pmatrix} U_{i1} - \bar{U}_i \\ \vdots \\ U_{iT} - \bar{U}_i \end{pmatrix}, \tag{6}$$

so that

$$Q_i = R_i\gamma + V_i.$$

Now we need orthogonality conditions. There is in fact considerable flexibility in setting up the W_i matrix. Here is one way to do it:

$$W_i = R_i' = \left((Z_{i1} - \bar{Z}_i) \quad \ldots \quad (Z_{iT} - \bar{Z}_i) \right), \tag{7}$$

so that

$$W_i V_i = \sum_{t=1}^{T} (Z_{it} - \bar{Z}_i)(U_{it} - \bar{U}_i).$$

The orthogonality condition $E(W_i V_i) = 0$ follows from

$$E(Z_{it}U_{is}) = 0 \qquad (s, t = 1, \ldots, T).$$

Note that it would not be enough to just have Z_{it} orthogonal to U_{it} for each t, because we need U_{it} to be orthogonal to $\bar{Z}_i$. That U_{it} is orthogonal to $(Z_{i1}, \ldots, Z_{iT})$ follows from (3)—U_{it} is a prediction error, where we are conditioning on $Z_{i1}, \ldots, Z_{iT}$ (and on A_i). This is where complete conditioning is used. It would not be enough in (3) to condition only on Z_{it} and A_i. We need to condition on Z at all dates and then impose the exclusion restriction that only Z_{it} matters for Y_{it}. If there were no exclusion restrictions, we would have no leverage for dealing with the unobserved variable A_i.

Note that with this choice for W_i, we have $L = K$, so that the number of orthogonality conditions matches the number of elements in γ. It will turn out that our system estimator is simply the least-squares fit of $(Y_{it} - \bar{Y}_i)$ on $(Z_{it} - \bar{Z}_i)$, pooling all NT observations:

$$\hat{\gamma} = \arg\min_a \sum_{i=1}^{N} \sum_{t=1}^{T} [(Y_{it} - \bar{Y}_i) - (Z_{it} - \bar{Z}_i)' a]^2. \tag{8}$$

This pooled least-squares estimator using deviations has shown up before—see Section 5 of Note 4 on "Stacking."

There is a way to set up the W_i matrix that maximizes the number of orthogonality conditions. Let

$$Z_i = \begin{pmatrix} Z_{i1} \\ \vdots \\ Z_{iT} \end{pmatrix}$$

and set

$$W_i = \begin{pmatrix} Z_i & 0 & \dots & 0 \\ 0 & Z_i & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & Z_i \end{pmatrix}. \tag{9}$$

Then

$$W_i V_i = \begin{pmatrix} Z_i(U_{i1} - \bar{U}_i) \\ \vdots \\ Z_i(U_{iT} - \bar{U}_i) \end{pmatrix} \tag{10}$$

and the number of orthogonality conditions is $L = T^2 \cdot K$. Now there are more orthogonality conditions than parameters to be estimated. Our estimator will allow for this by using a weight matrix.

Another way to eliminate A_i is to use first differences:

$$Y_{it} - Y_{i,t-1} = (Z_{it} - Z_{i,t-1})'\gamma + (U_{it} - U_{i,t-1}) \qquad (t = 2, \dots, T). \tag{11}$$

These $T - 1$ equations fit into the systems framework by setting

$$Q_i = \begin{pmatrix} Y_{i2} - Y_{i1} \\ \vdots \\ Y_{iT} - Y_{i,T-1} \end{pmatrix}, \quad R_i = \begin{pmatrix} (Z_{i2} - Z_{i1})' \\ \vdots \\ (Z_{iT} - Z_{i,T-1})' \end{pmatrix}, \quad V_i = \begin{pmatrix} U_{i2} - U_{i1} \\ \vdots \\ U_{iT} - U_{i,T-1} \end{pmatrix}. \tag{12}$$

For orthogonality conditions, we can set

$$W_i = R_i' = (\,(Z_{i2} - Z_{i1}) \quad \ldots \quad (Z_{iT} - Z_{i,T-1})\,)\,,$$

so that

$$W_i V_i = \sum_{t=2}^{T} (Z_{it} - Z_{i,t-1})(U_{it} - U_{i,t-1}).$$

This gives $L = K$, and our system estimator will turn out to be the least-squares fit of $(Y_{it} - Y_{i,t-1})$ on $(Z_{it} - Z_{i,t-1})$, pooling $N(T-1)$ observations:

$$\hat{\gamma} = \arg\min_{a} \sum_{i=1}^{N} \sum_{t=2}^{T} [(Y_{it} - Y_{i,t-1}) - (Z_{it} - Z_{i,t-1})'a]^2. \tag{13}$$

We can maximize the number of orthogonality conditions by using W_i as in (9) but with $T-1$ diagonal blocks, giving

$$W_i V_i = \begin{pmatrix} Z_i(U_{i2} - U_{i1}) \\ \vdots \\ Z_i(U_{iT} - U_{i,T-1}) \end{pmatrix} \tag{14}$$

and $T(T-1) \cdot K$ orthogonality conditions. This is less than the $T^2 \cdot K$ orthogonality conditions in (10), based on using deviations from time averages, but there is a linear dependence across the rows of $W_i V_i$ in (10):

$$Z_i(U_{i1} - \bar{U}_i) + \cdots + Z_i(U_{iT} - \bar{U}_i) = Z_i \sum_{t=1}^{T} (U_{it} - \bar{U}_i) = 0$$

since the deviations of U_{it} from the mean $\bar{U}_i$ sum to 0.

2.2 *Sequential Conditioning*

Consider the panel autoregression model from Note 5:

$$E(Y_{it} \,|\, Y_{i1}, \ldots, Y_{i,t-1}, A_i) = \lambda_t + \theta Y_{i,t-1} + A_i \qquad (t = 2, \ldots, T). \tag{15}$$

For example, Y_{it} is the log of earnings for individual i in year t, and we have earnings data for $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The variable A_i is not observed; it is an individual effect that is constant over time with variation only across individuals. The model for Y_{it} conditions on all the past values back to Y_{i1} and assumes that only $Y_{i,t-1}$ matters, once we control for A_i.

Define the prediction error

$$U_{it} = Y_{it} - E(Y_{it} \,|\, Y_i^{(t-1)}, A_i), \tag{16}$$

with

$$Y_i^{(s)} = \begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{is} \end{pmatrix}.$$

Then we can write the equations

$$Y_{it} = \lambda_t + \theta Y_{i,t-1} + A_i + U_{it} \qquad (t = 2, \ldots, T).$$

The sequential conditioning provides the following orthogonality conditions:

$$E(U_{it}) = 0, \quad E(Y_i^{t-1} U_{it}) = 0 \qquad (t = 2, \ldots, T). \tag{17}$$

We can eliminate A_i by taking first differences:

$$Y_{it} - Y_{i,t-1} = (\lambda_t - \lambda_{t-1}) + \theta(Y_{i,t-1} - Y_{i,t-2}) + (U_{it} - U_{i,t-1}) \qquad (t = 3, \ldots, T). \tag{18}$$

These $T - 2$ equations provide a system that maps into our framework in in (1). Let

$$Q_i = \begin{pmatrix} Y_{i3} - Y_{i2} \\ \vdots \\ Y_{iT} - Y_{i,T-1} \end{pmatrix}, \quad R_i = \begin{pmatrix} 1 & \ldots & 0 & Y_{i2} - Y_{i1} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \ldots & 1 & Y_{i,T-1} - Y_{i,T-2} \end{pmatrix}, \tag{19}$$

$$V_i = \begin{pmatrix} U_{i3} - U_{i2} \\ \vdots \\ U_{iT} - U_{i,T-1} \end{pmatrix}, \quad \gamma = \begin{pmatrix} \lambda_3 - \lambda_2 \\ \vdots \\ \lambda_T - \lambda_{T-1} \\ \theta \end{pmatrix}, \tag{20}$$

so that

$$Q_i = R_i\gamma + V_i.$$

For orthogonality conditions, we can use

$$W_i = \begin{pmatrix} I_{T-2} \\ W_{i2} \end{pmatrix}, \tag{21}$$

where I_{T-2} is the $(T-2)\times(T-2)$ identity matrix and

$$W_{i2} = \begin{pmatrix} Y_i^{(1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & Y_i^{(T-2)} \end{pmatrix}. \tag{22}$$

Then

$$W_iV = \begin{pmatrix} U_{i3} - U_{i2} \\ \vdots \\ U_{iT} - U_{i,T-1} \\ Y_i^{(1)}(U_{i3} - U_{i2}) \\ \vdots \\ Y_i^{(T-2)}(U_{iT} - U_{i,T-1}) \end{pmatrix}. \tag{23}$$

Note that the orthogonality conditions in (17) imply that

$$E[Y_i^{(t-2)}(U_{it} - U_{i,t-1})] = 0 \qquad (t = 3, \dots, T), \tag{24}$$

and so we have

$$E(W_iV_i) = 0.$$

3. MOMENT EQUATION

The estimator is based on the orthogonality condition $E(W_iV_i) = 0$. Multiply both sides of equation (1) by W_i and then average both sides of the equation over the sample:

$$W_iQ_i = W_iR_i\gamma + W_iV_i \tag{25}$$

$$\frac{1}{n}\sum_{i=1}^{n} W_iQ_i = \left(\frac{1}{n}\sum_{i=1}^{n} W_iR_i\right)\gamma + \frac{1}{n}\sum_{i=1}^{n} W_iV_i. \tag{26}$$

Let $S_{WQ} = \sum_i W_i Q_i / n$ and write (26) as

$$S_{WQ} = S_{WR}\gamma + S_{WV}. \tag{26$'$}$$

This *moment equation* is the key to obtaining our estimator and its properties. The law of large numbers implies that

$$S_{WQ} - S_{WR}\gamma = S_{WV} \xrightarrow{p} E(W_i V_i) = 0,$$

which suggests the *minimum distance* estimator

$$\hat{\gamma} = \arg\min_a ||S_{WQ} - S_{WR}a||^2. \tag{27}$$

The norm in the distance criterion corresponds to the inner product

$$\langle a, b \rangle = a'\hat{C}b,$$

where $\hat{C}$ is a positive definite, symmetric matrix. The only requirement on the weight matrix $\hat{C}$ for our large sample results is that it converge in probability to a nonrandom matrix C, which is also positive definite and symmetric. As in Section 4 of Note 4, the solution to this minimum norm problem is a *generalized least-squares* estimator:

$$\hat{\gamma} = (S'_{WR}\hat{C}S_{WR})^{-1} S'_{WR}\hat{C}S_{WQ}. \tag{28}$$

Note that S_{WQ} on the left-hand side of (26$'$) is $L \times 1$, so that (26$'$) provides L equations to solve for the K elements in $\hat{\gamma}$, when we replace S_{WV} by its limit of 0. If $L = K$, then S_{WR} is a square matrix and, assuming it is nonsingular, we can solve for $\hat{\gamma}$:

$$\hat{\gamma} = S_{WR}^{-1} S_{WQ}. \tag{29}$$

In this case, (28) reduces to (29) for any $\hat{C}$:

$$\hat{\gamma} = S_{WR}^{-1}\hat{C}^{-1}(S'_{WR})^{-1}S'_{WR}\hat{C}S_{WQ} = S_{WR}^{-1}S_{WQ} \quad \text{if } L = K$$

(using $(AB)^{-1} = B^{-1}A^{-1}$ when A and B are $K \times K$ nonsingular matrices). When $L = K$, we are in the *just-identified* case.

So we only need a weight matrix in the *over-identified* case $L > K$. If we set

$$\hat{D} = S_{WR}\hat{C}, \tag{30}$$

then we can write (28) in a simpler form:

$$\hat{\gamma} = (\hat{D}S_{WR})^{-1}\hat{D}S_{WQ}.$$

This suggests another approach, which applies a weight matrix directly to the moment equation:

$$\hat{D}S_{WQ} = (\hat{D}S_{WR})\gamma + \hat{D}S_{WV}. \tag{31}$$

We require that $\hat{D}$ be $K \times L$, with $\hat{D}S_{WR}$ nonsingular, and that $\hat{D}$ converge in probability to a nonrandom matrix D, with $DE(W_iR_i)$ nonsingular. Then

$$\hat{D}S_{WQ} - (\hat{D}S_{WR})\gamma = \hat{D}S_{WV} \xrightarrow{p} D \cdot 0 = 0,$$

which suggests the following estimator:

$$\hat{\gamma} = (\hat{D}S_{WR})^{-1}\hat{D}S_{WQ}. \tag{32}$$

This coincides with the minimum-distance estimator if $\hat{D} = S'_{WR}\hat{C}$, but the weight matrix $\hat{D}$ need not have this form.

The next section shows that $\hat{\gamma}$ in (32) is a consistent estimator for γ, and Section 5 develops a limit distribution. These results using the $\hat{D}$ weight matrix will also apply to the minimum-distance form of the estimator in (28), using the $\hat{C}$ weight matrix.

4. CONSISTENT ESTIMATION

We shall work with the $\hat{D}$ form of $\hat{\gamma}$ in (32). Assume that $\hat{D} \xrightarrow{p} D$, a $K \times L$ nonrandom matrix, and that $DE(W_iR_i)$ is nonsingular. The proof is similar to the one for least squares in Note 8, Section 1.

Claim 1. $\hat{\gamma} \xrightarrow{p} \gamma$ as $n \to \infty$.

Proof. By the law of large numbers and Slutsky (i),

$$\hat{\gamma} \xrightarrow{p} [DE(W_i R_i)]^{-1} DE(W_i Q_i).$$

Multiply (25) by D and take expectations:

$$DW_i Q_i = DW_i R_i \gamma + DW_i V_i,$$
$$DE(W_i Q_i) = [DE(W_i R_i)]\gamma + 0.$$

Substitute into the probability limit for $\hat{\gamma}$ to obtain $\hat{\gamma} \xrightarrow{p} \gamma$. $\diamond$

5. LIMIT DISTRIBUTION

The limit distribution and the proof are similar to the results for least squares in Note 8, Section 2.

Claim 2. $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} \mathcal{N}(\alpha \Sigma \alpha')$ where

$$\alpha = [DE(W_i R_i)]^{-1} D, \quad \Sigma = E(W_i V_i V_i' W_i').$$

Proof. Substituting (26′) into (32) gives

$$\hat{\gamma} = (\hat{D} S_{WR})^{-1} \hat{D}(S_{WR}\gamma + S_{WV})$$
$$= \gamma + (\hat{D} S_{WR})^{-1} \hat{D} S_{WV},$$

so that

$$\sqrt{n}(\hat{\gamma} - \gamma) = (\hat{D} S_{WR})^{-1} \hat{D} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_i V_i.$$

Define $G_i = W_i V_i$. Since G_i is a function of (Q_i, R_i, W_i), random sampling implies that the G_i are independent and identically distributed. In addition,

$$E(G_i) = 0, \quad \mathrm{Cov}(G_i) = E(G_i G_i') = \Sigma,$$

and so the central limit theorem implies that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_i \xrightarrow{d} \mathcal{N}(0,\Sigma).$$

As in our proof that $\hat{\gamma}$ is a consistent estimator of γ, the law of large numbers and Slutsky (i) imply that

$$(\hat{D}S_{WR})^{-1}\hat{D} \xrightarrow{p} [DE(W_iR_i)]^{-1}D = \alpha.$$

Then Slutsky (ii) implies that

$$\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} \alpha\cdot\mathcal{N}(0,\Sigma) = N(0,\alpha\Sigma\alpha'). \quad \diamond$$

6. CONFIDENCE INTERVAL

Define $\Lambda = \alpha\Sigma\alpha'$, so that Claim 2 gives

$$\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} \mathcal{N}(0,\Lambda).$$

We are going to follow the argument for least squares in Note 8, Section 3 to use the limit distribution to construct a confidence interval. In order to obtain a consistent estimate of Λ, let $\hat{V}_i = Q_i - R_i\hat{\gamma}$ and define

$$\hat{\Lambda} = \hat{\alpha}\hat{\Sigma}\hat{\alpha}',$$

with

$$\hat{\alpha} = (\hat{D}S_{WR})^{-1}\hat{D}, \quad \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} W_i\hat{V}_i\hat{V}_i'W_i'.$$

We can use the law of large numbers and Slutsky (i) to show that

$$\hat{\Lambda} \xrightarrow{d} \Lambda.$$

As in Section 4 of Note 7, we shall obtain a confidence interval for a linear combination of the coefficients:

$$l'\gamma = \sum_{j=1}^{K} l_j\gamma_j.$$

Slutsky (i) implies that

$$(l'\hat{\Lambda}l)^{1/2} \xrightarrow{p} (l'\Lambda l)^{1/2},$$

and Slutsky (ii) implies that

$$\frac{l'[\sqrt{n}(\hat{\gamma}-\gamma)]}{(l'\hat{\Lambda}l)^{1/2}} \xrightarrow{d} \frac{1}{(l'\Lambda l)^{1/2}}\mathcal{N}(0, l'\Lambda l) = \mathcal{N}(0,1).$$

Define the standard error as

$$\mathrm{SE} = (l'\hat{\Lambda}l/n)^{1/2}.$$

We have established

Claim 3.

$$\frac{l'(\hat{\gamma}-\gamma)}{\mathrm{SE}} \xrightarrow{d} \mathcal{N}(0,1).$$

The ratio $l'(\hat{\gamma}-\gamma)/\mathrm{SE}$ is an asymptotic pivot for $l'\gamma$. It depends upon the unknown parameters only through $l'\gamma$, and it has a known limit distribution. This leads to a confidence interval for $l'\gamma$. Claim 3 gives

$$\lim_{n\to\infty} \mathrm{Prob}(-1.96 \le \frac{l'\gamma - l'\hat{\gamma}}{\mathrm{SE}} \le 1.96) = .95,$$

and so

$$\lim_{n\to\infty} \mathrm{Prob}(l'\hat{\gamma} - 1.96\cdot\mathrm{SE} \le l'\gamma \le l'\hat{\gamma} + 1.96\cdot\mathrm{SE}) = .95,$$

or

$$\lim_{n\to\infty} \mathrm{Prob}(l'\gamma \in [l'\hat{\gamma} \pm 1.96\cdot\mathrm{SE}]) = .95.$$

7. CONFIDENCE ELLIPSE

We can follow the procedure for the normal linear model in Note 7, Section 5. Let L be $h\times K$ so that $L\gamma$ is $h\times 1$. Then Slutsky (ii) implies that

$$L\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} L\cdot\mathcal{N}(0,\Lambda) = \mathcal{N}(0, L\Lambda L')$$

and

$$S_n \equiv (L\hat{\Lambda}L')^{-1/2}L\sqrt{n}(\hat{\gamma}-\gamma) \xrightarrow{d} \mathcal{N}(0, I_h). \tag{33}$$

Since S_n converges in distribution to $\mathcal{N}(0, I_h)$, the sum of squares $S'S$ converges in distribution to $\text{Chi}^2(h)$. This follows from another part of the Slutsky theorem, which is also known as the *continuous mapping theorem*:

Slutsky Theorem. (iii) Let S_n be a sequence of $h \times 1$ random variables that converges in distribution; let S be a random variable whose distribution is the limit distribution of S_n: $S_n \xrightarrow{d} S$. Then if $g : \mathcal{R}^h \to \mathcal{R}^m$ is a continuous function, $g(S_n) \xrightarrow{d} g(S)$.

Define

$$\hat{\text{Var}}(L\hat{\gamma}) = L\hat{\Lambda}L'/n.$$

Claim 4. $(L\hat{\gamma} - L\gamma)'[\hat{\text{Var}}(L\hat{\gamma})]^{-1}(L\hat{\gamma} - L\gamma) \xrightarrow{d} \text{Chi}^2(h)$.

Proof.

$$S_n'S_n = (L\hat{\gamma} - L\gamma)'(L\hat{\Lambda}L'/n)^{-1}(L\hat{\gamma} - L\gamma);$$

(33) and Slutsky (iii) imply that

$$S_n'S_n \xrightarrow{d} \text{Chi}^2(h). \quad \diamond$$

Claim 4 provides an asymptotic pivot for $L\gamma$, because

$$(L\hat{\gamma} - L\gamma)'[\hat{\text{Var}}(L\hat{\gamma})]^{-1}(L\hat{\gamma} - L\gamma) \tag{34}$$

depends upon the unknown parameters only through $L\gamma$, and it has a known distribution. This leads to a confidence region for $L\gamma$.

For example, suppose that $h = 2$ with

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \end{pmatrix} \quad \text{and} \quad L\gamma = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}.$$

The Chi^2 distribution is available in tables and computer programs. We have

$$\mathrm{Prob}(\mathrm{Chi}^2(2) > 5.99) = .05.$$

Then Claim 4 gives

$$\lim_{n\to\infty} \mathrm{Prob}([\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} - \begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{pmatrix}]'[\hat{\mathrm{Var}}\begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{pmatrix}]^{-1}[\begin{pmatrix}\gamma_1\\ \gamma_2\end{pmatrix} - \begin{pmatrix}\hat{\gamma}_1\\ \hat{\gamma}_2\end{pmatrix}] \le 5.99) = .95. \tag{35}$$

The confidence region consists of the values for (γ_1, γ_2) that satisfy the inequality in (35). This gives the interior of an ellipse which is centered at $(\hat{\gamma}_1, \hat{\gamma}_2)$.

8. SERIAL CORRELATION AND THE HOMOSKEDASTIC CASE

So far our basic assumptions have been random sampling and the the orthogonality condition $E(W_i V_i) = 0$. In particular, we have *not* assumed that the conditional expectation of V_i given W_i is zero, or that the conditional covariance matrix of V_i given W_i is constant. Now we are going to see how the asymptotic inference simplifies if we do make these additional assumptions. This will make it easier to see how the asymptotic inference deals with correlation across the components of the error vector V_i, as would be caused by serial correlation in longitudinal panel data.

So assume that

$$\text{(i) } E(V_i \,|\, W_i) = 0 \quad \text{and} \quad \text{(ii) } \mathrm{Cov}(V_i \,|\, W_i) = E(V_i V_i' \,|\, W_i) = \Omega.$$

Here Ω is a constant $T \times T$ matrix. This implies that Σ depends only upon second moments instead of fourth moments:

$$\begin{aligned}\Sigma &= E[E(W_i V_i V_i' W_i' \,|\, W_i)] \\ &= E[W_i E(V_i V_i' \,|\, W_i) W_i'] \\ &= E(W_i \Omega W_i').\end{aligned}$$

Because $\Omega = E(V_i V_i')$, a consistent estimate of Ω is

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} \hat{V}_i \hat{V}_i',$$

and a consistent estimate of Σ is

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} W_i \hat{\Omega} W_i'.$$

In the panel data application in Section 2.1,

$$V_i = \begin{pmatrix} U_{i1} - \bar{U}_i \\ \vdots \\ U_{iT} - \bar{U}_i \end{pmatrix},$$

and the (s,t) element of Ω is

$$\Omega_{st} = E(V_{is} V_{it}) = E[(U_{is} - \bar{U}_i)(U_{it} - \bar{U}_i)] \qquad (s, t = 1, \ldots, T).$$

Serial correlation in the errors U_{it} shows up in $E(U_{is} U_{it}) \neq 0$. Even if there is no such serial correlation in the U's, the transformed errors V_{it} will have serial correlation because U_{it} is correlated with $\bar{U}_i$. Our estimator for Ω deals with this by using the sample covariance of the estimated errors:

$$\hat{\Omega}_{st} = \frac{1}{n} \sum_{i=1}^{n} \hat{V}_{is} \hat{V}_{it}.$$