SYSTEMS OF EQUATIONS: SYSTEM OLS

Econometric Analysis of Cross Section and Panel Data, 2e
MIT Press
Jeffrey M. Wooldridge

1. Examples of “Systems” of Equations
2. System OLS Estimation
3. Extensions

1. EXAMPLES OF "SYSTEMS" OF EQUATIONS

- Carry along two examples: Seemingly Unrelated Regressions (SUR) and panel data.

- We assume random sampling of units from a well-defined population. We may sample several different response variables along with explanatory variables (SUR), or sample different time periods on the same response and explanatory variables (panel data).

- In the panel data case, we assume a small number of time periods; to apply standard limit theorems (law of large numbers, central limit theorem), we can use results for independent, indentically distributed observations.

The SUR Case

- A G-equation SUR system is written in the population as

$$\begin{aligned} y_1 &= \mathbf{x}_1\boldsymbol{\beta}_1 + u_1 \\ y_2 &= \mathbf{x}_2\boldsymbol{\beta}_2 + u_2 \\ &\vdots \\ y_G &= \mathbf{x}_G\boldsymbol{\beta}_G + u_G \end{aligned} \tag{1.1}$$

where y_g is a response variable, $g = 1, \ldots, G$. The explanatory variables, $\mathbf{x}_g$, can be different across equations. For now, think if the $\boldsymbol{\beta}_g$ as being unrestricted across equations.

• What might we assume about exogeneity of the explanatory variables? In terms of conditional means, two possibilities:

$$E(u_g|\mathbf{x}_g) = 0,\ g = 1,\ldots,G, \tag{1.2}$$

which means

$$E(y_g|\mathbf{x}_g) = \mathbf{x}_g\boldsymbol{\beta}_g,\ g = 1,\ldots,G. \tag{1.3}$$

• We could instead use the weaker condition

$$E(\mathbf{x}_g'u_g) = \mathbf{0},\ g = 1,\ldots,G. \tag{1.4}$$

Key point is that neither (1.2) nor (1.4) restricts the relationship between $\mathbf{x}_h$ and u_g for $g \neq h$.

• A stronger assumption, implicitly or explicitly maintained by most SUR analyses, is

$$E(u_g|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g, \ldots, \mathbf{x}_G) = 0, \; g = 1, \ldots, G, \tag{1.5}$$

which implies

$$E(y_g|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g, \ldots, \mathbf{x}_G) = E(y_g|\mathbf{x}_g) = \mathbf{x}_g\boldsymbol{\beta}_g, \; g = 1, \ldots, G. \tag{1.6}$$

• This means that, if $\mathbf{x}_h$ for $h \neq g$ includes elements not in $\mathbf{x}_g$, then those elements of $\mathbf{x}_h$ are assumed to have no partial effect on the expected value of y_g once $\mathbf{x}_g$ is controlled for.

- Unless $\mathbf{x}_g = \mathbf{x}_h$ for all g and h,

$$E(u_g|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_g, \ldots, \mathbf{x}_G) = 0$$

imposes substantive exclusion restrictions. Treating the explanatory variables as fixed in repeated samples is operationally the same as this assumption (at least in terms of obtaining statistical properties).

- If $\mathbf{x}_g = \mathbf{x}_h$ for all g and h, there is no difference between (1.2) and (1.5).

EXAMPLE: Suppose each worker in a population receives three kinds of compensation: wage, pension, and health:

$$\begin{aligned} wage &= \beta_{10} + \beta_{11}educ + \beta_{12}tenure + \beta_{13}age + \beta_{14}union + u_1 \\ pension &= \beta_{20} + \beta_{21}educ + \beta_{22}tenure + \beta_{23}age + \beta_{24}union + u_2 \\ health &= \beta_{30} + \beta_{31}educ + \beta_{32}tenure + \beta_{33}age + \beta_{34}union + u_3 \end{aligned} \tag{1.7}$$

Then $G = 3$, and our random sample consists of workers from the specified population. We have three response variables in our population model.

• In some applications, especially to consumer and firm theory, the coefficients are restricted across equations (later).

The Panel Data Case

- Now suppose that, for each unit in the population, we have a (usually short) time series, $\{(\mathbf{x}_t, y_t) : t = 1, \ldots, T\}$. A linear panel data model is

$$y_t = \mathbf{x}_t \boldsymbol{\beta} + u_t, \; t = 1, \ldots, T. \tag{1.8}$$

- In this setup, there is a single response variable, y_t, that we observe in several time periods.
- Having the same $\boldsymbol{\beta}$ for all t is not restrictive because $\mathbf{x}_t$ can be very general. It can (and usually should) include time-period dummies to allow a different intercept in each period. It can include interactions between time-period dummies and other variables to allow partial effects to change over time.

- We can write (1.8) in a way similar to a SUR system:

$$
\begin{aligned}
y_1 &= \mathbf{x}_1\boldsymbol{\beta} + u_1 \\
y_2 &= \mathbf{x}_2\boldsymbol{\beta} + u_2 \\
&\vdots \\
y_T &= \mathbf{x}_T\boldsymbol{\beta} + u_T
\end{aligned}
\tag{1.9}
$$

- Suggests that, even though the applications are very different, a common statistical framework can be used for SUR and panel data.
- $\mathbf{x}_t$ has the same dimension for all t. In SUR, the dimension of the covariates generally changes across equation.

EXAMPLE: Suppose that for the years 2000 through 2003 an annual family saving equation is

$$
\begin{aligned}
sav_t = {} & \beta_0 + \beta_1 d01_t + \beta_2 d02_t + \beta_3 d03_t \\
& + \beta_4 inc_t + \beta_5 size_t + \beta_6 educ_t + \beta_7 e401k_t \\
& + \beta_8 d01_t \cdot e401k_t + \beta_9 d02_t \cdot e401k_t + \beta_{10} d03_t \cdot e401k_t + u_t
\end{aligned}
\tag{1.10}
$$

- A variable such as $educ_t$, education of the household head, might not change over time for many of the units. Some variables, such as gender, do not change over time for any units in the population.
- The general model allows variables that change only across time (such as year dummies), only across unit (such as gender), and across unit and time (typically, the most interesting variables).

EXAMPLE: A distributed lag model for county-level poverty rates might look like

$$\begin{aligned} poverty_t &= \alpha_t + \delta_0 welfare_t + \delta_1 welfare_{t-1} + \delta_2 welfare_{t-2} \\ &\quad + \gamma_1 educ_t + \gamma_2 perc\_children_t + \gamma_3 perc\_elderly_t + u_t, \end{aligned} \tag{1.11}$$

where $welfare_t$ is per-capital welfare spending. In deciding how to estimate a model such as this, we must be very careful in stating exogeneity assumptions for the explanatory variables.

- Education and age distribution variables are just a couple of possible controls. The α_t are different period intercepts.
- The long run propensity is $\delta_0 + \delta_1 + \delta_2$, which is often easier to estimate than the individual δ_j.

- An attractive assumption in the general equation $y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t$ is **contemporaneous exogeneity**:

$$E(u_t|\mathbf{x}_t) = 0, t = 1, \ldots, T, \tag{1.12}$$

which implies

$$E(y_t|\mathbf{x}_t) = \mathbf{x}_t\boldsymbol{\beta}, \; t = 1, \ldots, T. \tag{1.13}$$

- Rules out omitted variables, measurement error, and so on, but it does not restrict correlation between $\mathbf{x}_s$ and u_t for $s \neq t$.
- Stated as a zero conditional mean, it also implies correct functional form of $E(y_t|\mathbf{x}_t)$.

• An assumption that is stronger than (1.12), but can be justified when we assume the dynamics (with respect to the elements of $\mathbf{x}_t$) are completely specified, is **sequential exogeneity**:

$$E(u_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = 0,\ t = 1, \ldots, T, \tag{1.14}$$

• As we will see, (1.14) can be applied to distributed lag models and models with lagged dependent variables.

• (1.14) implies

$$E(y_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) = E(y_t|\mathbf{x}_t) = \mathbf{x}_t\boldsymbol{\beta}, \tag{1.15}$$

which makes it clear that enough lags of whatever have been included in $\mathbf{x}_t$ so that no further lags are needed.

- An even stronger assumption, both technically and practically, is **strict exogeneity**:

$$E(u_t|\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{x}_{t+1},\ldots,\mathbf{x}_T) = 0,\ t = 1,\ldots,T. \tag{1.17}$$

This, of course, implies that $\mathbf{x}_s$ and u_t are uncorrelated for all s and t, including $s < t$, $s = t$, and $s > t$.

- In the context of $y_t = \mathbf{x}_t\boldsymbol{\beta} + u_t$, (1.17) is the same as

$$E(y_t|\mathbf{x}_1,\ldots,\mathbf{x}_{t-1},\mathbf{x}_t,\mathbf{x}_{t+1},\ldots,\mathbf{x}_T) = E(y_t|\mathbf{x}_t) = \mathbf{x}_t\boldsymbol{\beta}, \tag{1.18}$$

where the first equality is the critical one. It says that, given whatever is in $\mathbf{x}_t$, the covariates from other time periods do not help explain y_t.

- The main concern with (1.18) is that it rules out feedback from shocks to y at time t to future outcomes on $\mathbf{x}$. In particular, u_t and $\mathbf{x}_{t+1}$ cannot be correlated.
- An individual might become married or divorced in year $t + 1$ depending on the shock in time t. A school might adjust class size next year based on shocks to student peformance this year.

EXAMPLE: A firm-level production function might look like

$$\log(Q_t) = \alpha_t + \beta_1 \log(K_t) + \beta_2 \log(L_t) + \beta_3 \log(M_t) + u_t, \tag{1.19}$$

where u_t contains productivity and other shocks. The strict exogeneity assumption rules out the possibility that firms adjust inputs (capital, labor, materials) at, say, time $t + 1$ in reaction to shocks at time t.

• This example shows why maintaining fixed regressors in economic applications with panel data is a non-starter.

• We may or may not think strict or contemporaneous exogeneity hold for the production function, but at least the question makes sense when we view $\mathbf{x}_t$ as a random outcome along with u_t.

• In the poverty/welfare spending distributed lag example, we might be worried about "feedback effects." For example, it might make sense to think of welfare payments being generated by something like

$$welfare_{t+1} = \eta_t + \lambda_1 poverty_t + r_{t+1}, \tag{1.20}$$

which would violate strict exogeneity.

• Any model with a lagged dependent variable *must* violate strict exogeneity. For example, for a population of students,

$$score_t = \rho_1 score_{t-1} + \gamma_1 classize_t + \gamma_2 faminc_t + u_t$$

• The reason for putting in the lagged test score is to control for the possibility that classroom assignment depends on prior performance.

- In general, write a model with a single lag as

$$
\begin{aligned}
y_t &= \rho_1 y_{t-1} + \mathbf{z}_t \boldsymbol{\gamma} + u_t, \ t = 1, 2, \ldots, T \\
E(u_t | y_{t-1}, \mathbf{z}_t) &= 0
\end{aligned}
$$

Then $\mathbf{x}_t = (y_{t-1}, \mathbf{z}_t)$ and so

$$
(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = (y_0, y_1, \ldots, y_{T-1}, \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_T)
$$

Thefore, for $t < T$,

$$
E(y_t | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T) = y_t \neq E(y_t | \mathbf{x}_t)
$$

- As with finite distributed lage models, for models with lagged *y* we often want sequential exogeneity to hold – and that may or may not be true. But strict exogeneity cannot hold.
- We discuss dynamic models in more detail later.

2. SYSTEM OLS ESTIMATION

2.1. Consistency

• First estimation method uses OLS on the system of equations. We will be interested in what this entails for the SUR and panel data cases.

• It is notationally useful to use an *i* subscript when writing the system, to distinguish between unit-specific observations and full data matrices. We write the model for a random draw *i* as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i \tag{2.1}$$

where $\mathbf{y}_i$ is $G \times 1$, $\mathbf{X}_i$ is $G \times K$, $\boldsymbol{\beta}$ is $K \times 1$, and $\mathbf{u}_i$ is $G \times 1$. The observed data is $\{(\mathbf{X}_i, \mathbf{y}_i) : i = 1, \ldots, N\}$, where N is the sample size.

• We want to estimate the population parameter vector $\boldsymbol{\beta}$.

SUR

In the SUR case, $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iG})'$, $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iG})'$,

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{i2} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} & \vdots \\ \vdots & \vdots & \mathbf{0} & \mathbf{x}_{i,G-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_{iG} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{G-1} \\ \boldsymbol{\beta}_G \end{pmatrix}. \tag{2.2}$$

If $\mathbf{x}_{ig}$ is $1 \times K_g$, define $K = K_1 + K_2 + \ldots + K_G$, and then $\mathbf{X}_i$ is $G \times K$ and $\boldsymbol{\beta}$ is $K \times 1$.

Panel Data

In the panel data case, $G = T$ (number of time periods), $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{iT})'$, $\mathbf{u}_i = (u_{i1}, u_{i2}, \ldots, u_{iT})'$, and

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{iT} \end{pmatrix}, \tag{2.3}$$

where each $\mathbf{x}_{it}$ is $1 \times K$. (Contrast the SUR case, where the dimension of the covariates can be different across equations. Remember, though, that $\mathbf{x}_{it}$ can be chosen quite flexibly.)

Assumptions for System OLS (SOLS)

Assumption SOLS.1:

$$E(\mathbf{X}_i'\mathbf{u}_i) = \mathbf{0}. \ \square \tag{2.4}$$

- This is the weakest possible assumption without moving into instrumental variables territory.
- Clearly (2.4) is implied by

$$E(\mathbf{u}_i|\mathbf{X}_i) = \mathbf{0}, \tag{2.5}$$

almost the strongest assumption we could make (short of independence between $\mathbf{u}_i$ and $\mathbf{X}_i$).

- In the SUR case,

$$\mathbf{X}_i'\mathbf{u}_i = \begin{pmatrix} \mathbf{x}_{i1}'u_{i1} \\ \mathbf{x}_{i2}'u_{i2} \\ \vdots \\ \mathbf{x}_{iG}'u_{iG} \end{pmatrix} \tag{2.6}$$

and so SOLS.1 is equivalent to

$$E(\mathbf{x}_{ig}'u_{ig}) = \mathbf{0},\ g = 1, \ldots, G. \tag{2.7}$$

- SOLS.1 does not restrict relationship between u_{ig} and covariates in other equations.

- In the panel data case,

$$\mathbf{X}_i'\mathbf{u}_i = \sum_{t=1}^{T} \mathbf{x}_{it}'u_{it} \tag{2.8}$$

and so SOLS.1 is the same as

$$\sum_{t=1}^{T} E(\mathbf{x}_{it}'u_{it}) = \mathbf{0}. \tag{2.9}$$

- It is unlikely we would assume (2.9) without just assuming

$$E(\mathbf{x}_{it}'u_{it}) = \mathbf{0}, t = 1, \ldots, T. \tag{2.10}$$

- SOLS.1 for panel data is contemporaneous exogeneity.

- The rank condition for system OLS is

Assumption SOLS.2:

$$rank\, E(\mathbf{X}_i'\mathbf{X}_i) = K. \quad \square \tag{2.11}$$

SUR

- In the SUR case,

$$\mathbf{X}_i'\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1}'\mathbf{x}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{i2}'\mathbf{x}_{i2} & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{x}_{iG}'\mathbf{x}_{iG} \end{pmatrix} \tag{2.12}$$

- SOLS.2 holds if and only if

$$rank\, E(\mathbf{x}_{ig}'\mathbf{x}_{ig}) = K_g,\, g = 1, \ldots, G, \tag{2.13}$$

which simply says that the single-equation OLS rank condition (OLS.2) holds for each equation.

Panel Data

$$\mathbf{X}_i'\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{i1}' & \mathbf{x}_{i2}' & \cdots & \mathbf{x}_{iT}' \end{pmatrix} \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \\ \vdots \\ \mathbf{x}_{iT} \end{pmatrix} = \sum_{t=1}^{T} \mathbf{x}_{it}'\mathbf{x}_{it}. \tag{2.14}$$

- SOLS.2 holds if and only if

$$rank\left(\sum_{t=1}^{T} E(\mathbf{x}_{it}'\mathbf{x}_{it})\right) = K. \tag{2.15}$$

- There are cases (later on) where, for each t, the rank condition would not hold, but it does hold averaged across t.

Population Orthogonality Conditions

$$\begin{aligned} E[\mathbf{X}_i'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})] &= \mathbf{0} \text{ by SOLS.1} \\ E(\mathbf{X}_i'\mathbf{X}_i)\boldsymbol{\beta} &= E(\mathbf{X}_i'\mathbf{y}_i) \\ \boldsymbol{\beta} &= [E(\mathbf{X}_i'\mathbf{X}_i)]^{-1}E(\mathbf{X}_i'\mathbf{y}_i) \text{ by SOLS.2} \end{aligned} \tag{2.16}$$

• The SOLS estimator looks just like the single-equations OLS estimator, but $\mathbf{X}_i$ is a matrix and $\mathbf{y}_i$ is a vector:

$$\hat{\boldsymbol{\beta}}_{SOLS} = \left(N^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(N^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{y}_i\right), \tag{2.17}$$

which can be written as $\hat{\boldsymbol{\beta}}_{SOLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ where $\mathbf{X}$ is $NG \times K$ and $\mathbf{Y}$ is $NG \times 1$.

THEOREM: Under SOLS.1 and SOLS.2, OLS on a random sample is consistent:

$$\text{plim}_{N\to\infty}(\hat{\boldsymbol{\beta}}_{SOLS}) = \boldsymbol{\beta}. \tag{2.18}$$

- In the SUR case, it is obvious that we think of G fixed equations and collect more and more data on units (so N increases).
- In the panel data case, it is less obvious how to think of the asymptotics. Here we assume fixed T (number of time periods) with N growing. Reflects most micro panel data sets: many units (families, firms), few time periods per unit.

• The fixed T scenario is very convenient: it means the limit theorems are for random sampling, the stuff of an introductory probability and statistics course.

• Further, the results allow *any* kind of time series dependence in the explanatory variables or errors.

• The case $T \to \infty$ is more in the realm of time series econometrics and would require restricting the time series correlation to get unified results.

- What is the SOLS estimator in the SUR case? Use

$$\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{X}_i = \begin{pmatrix} \sum_{i=1}^{N} \mathbf{x}_{i1}' \mathbf{x}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sum_{i=1}^{N} \mathbf{x}_{i2}' \mathbf{x}_{i2} & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \sum_{i=1}^{N} \mathbf{x}_{iG}' \mathbf{x}_{iG} \end{pmatrix}$$

$$\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{y}_i = \begin{pmatrix} \sum_{i=1}^{N} \mathbf{x}_{i1}' y_{i1} \\ \sum_{i=1}^{N} \mathbf{x}_{i2}' y_{i2} \\ \vdots \\ \sum_{i=1}^{N} \mathbf{x}_{iG}' y_{iG} \end{pmatrix}, \tag{2.19}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \sum_{i=1}^{N} \mathbf{x}_{i1}'\mathbf{x}_{i1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sum_{i=1}^{N} \mathbf{x}_{i2}'\mathbf{x}_{i2} & \mathbf{0} & \vdots \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \mathbf{0} & \sum_{i=1}^{N} \mathbf{x}_{iG}'\mathbf{x}_{iG} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{N} \mathbf{x}_{i1}'y_{i1} \\ \sum_{i=1}^{N} \mathbf{x}_{i2}'y_{i2} \\ \vdots \\ \sum_{i=1}^{N} \mathbf{x}_{iG}'y_{iG} \end{pmatrix}$$

$$= \begin{pmatrix} \left(\sum_{i=1}^{N} \mathbf{x}_{i1}'\mathbf{x}_{i1}\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_{i1}'y_{i1} \\ \left(\sum_{i=1}^{N} \mathbf{x}_{i2}'\mathbf{x}_{i2}\right)^{-1} \sum_{i=1}^{N} \mathbf{x}_{i2}'y_{i2} \\ \vdots \\ (\sum_{i=1}^{N} \mathbf{x}_{iG}'\mathbf{x}_{iG})^{-1} \sum_{i=1}^{N} \mathbf{x}_{iG}'y_{iG} \end{pmatrix} \equiv \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_G \end{pmatrix}$$

- Each $\hat{\boldsymbol{\beta}}_g$ is just the OLS estimator on equation g.
- In this case, system OLS is **ordinary least squares equation-by-equation**.
- Of course, we already know that the conditions sufficient of consistency of OLS on equation g with random sampling are $E(\mathbf{x}_{ig}'u_{ig}) = \mathbf{0}$, $rank\, E(\mathbf{x}_{ig}'\mathbf{x}_{ig}) = K_g$.

• For panel data,

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}' \mathbf{x}_{it} \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x}_{it}' y_{it} \right), \tag{2.20}$$

which we call the **pooled OLS estimator**.

• System OLS is not unbiased under SOLS.1 and SOLS.2. It is (under some assumptions about moment conditions) if we use

$$E(\mathbf{u}_i | \mathbf{X}_i) = \mathbf{0}, \tag{2.21}$$

but this, as discussed before, is very strong.

2.2. Asymptotic Normality and Inference

- The proof of asymptotic normality is similar to the single-equation case, but we have to be more careful with the linear algebra. Write

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(N^{-1} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{X}_i \right)^{-1} \left(N^{-1/2} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{u}_i \right). \tag{2.22}$$

- By the CLT for i.i.d. random vectors,

$$N^{-1/2} \sum_{i=1}^{N} \mathbf{X}_i' \mathbf{u}_i \xrightarrow{d} Normal(\mathbf{0}, \mathbf{B}) \tag{2.23}$$

$$\mathbf{B} = Var(\mathbf{X}_i' \mathbf{u}_i) = E(\mathbf{X}_i' \mathbf{u}_i \mathbf{u}_i' \mathbf{X}_i). \tag{2.24}$$

• Define

$$\mathbf{A} = E(\mathbf{X}_i'\mathbf{X}_i). \tag{2.25}$$

$$\begin{aligned} N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= \mathbf{A}^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{u}_i\right) \\ &\quad + \left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right)^{-1} - \mathbf{A}^{-1}\right]\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{u}_i\right) \\ &= \mathbf{A}^{-1}\left(N^{-1/2}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{u}_i\right) + o_p(1)\cdot O_p(1). \end{aligned} \tag{2.26}$$

- So

$$N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} Normal(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}). \tag{2.27}$$

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i \quad \text{(a } K \times K \text{ matrix)} \tag{2.28}$$

- The SOLS residuals are

$$\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}} \quad \text{(each is } G \times 1\text{)} \tag{2.29}$$

• A fully robust estimator of $\mathbf{B}$ – that is, an estimator valid under SOLS.1 and SOLS.2, without any second moment assumptions on $\mathbf{u}_i$ – is

$$\hat{\mathbf{B}} = (N-K)^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{X}_i. \tag{2.30}$$

• The degrees-of-freedom adjustment is common but unecessary with large N. The resulting sandwich estimator is

$$\widehat{Avar}(\hat{\boldsymbol{\beta}}) = \frac{N}{(N-K)}\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{X}_i\right)\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{X}_i\right)^{-1}. \tag{2.31}$$

- In the SUR case, (2.31) allows arbitrary covariances in the errors across the different equations, different variances, and allows all conditional variances and covariances to be unknown functions of $\mathbf{X}_i$.
- In the panel data case, (2.31) allows the variances, $Var(u_t)$, and covariances, $Cov(u_t, u_s)$, to be unrestricted. Thus, it is robust to arbitrary heteroskedasticity as a function of time and to any serial correlation pattern in the errors, $\{u_{it} : t = 1, \ldots, T\}$. In addition, it is robust to the conditional variances and covariances being any function of $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$.

Testing

• Obtain asymptotic standard errors from (2.31) to construct large-sample t statistics (which converge to standard normal under the null as $N \to \infty$) and confidence intervals.

• Consider hypotheses

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}, \tag{2.32}$$

where $\mathbf{R}$ is $Q \times K$, $\mathbf{r}$ is $Q \times 1$, $Q \leq K$. Most convenient is the **Wald statistic**:

$$\mathcal{W} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'(\mathbf{R}\hat{\mathbf{V}}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \overset{a}{\sim} \chi^2_Q \tag{2.33}$$

under H_0, where $\hat{\mathbf{V}}$ is given in (2.31).

3. EXTENSIONS

- Later we will cover SUR systems where the parameters are restricted across equations. Then SOLS is not OLS equation-by-equation.
- We can combine SUR and panel data models. For example, for a random draw i, write

$$y_{itg} = \mathbf{x}_{itg}\boldsymbol{\beta}_g + u_{itg},\ \ g = 1,\ldots,G;\ t = 1,\ldots,T. \tag{3.1}$$

For each (i,t), we have a G-equation SUR system:

$$\mathbf{y}_{it} = \mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{u}_{it} \tag{3.2}$$

where $\mathbf{X}_{it}$ is $G \times K$ where $K = K_1 + \ldots + K_G$, as before.

- Stack the data for a random draw i to get the $TG \times 1$ vector $\mathbf{y}_i$:

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{i1} \\ \vdots \\ \mathbf{y}_{it} \\ \vdots \\ \mathbf{y}_{iT} \end{pmatrix} = \begin{pmatrix} y_{i11} \\ y_{i12} \\ \vdots \\ y_{i1G} \\ \vdots \\ y_{iT1} \\ y_{iT2} \\ \vdots \\ y_{iTG} \end{pmatrix} \tag{3.3}$$

and the $TG \times K$ matix $\mathbf{X}_i$.

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_{i1} \\ \vdots \\ \mathbf{X}_{it} \\ \vdots \\ \mathbf{X}_{iT} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{i11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{i12} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{i1G} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_{iT1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{iT2} & & \vdots \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{x}_{iTG} \end{pmatrix} \tag{3.4}$$

- Exercise: Characterize the SOLS estimator in this case.