

系统估计及其应用

司继春

2026年9月



为什么需要系统估计？

- 普通最小二乘通常处理单方程： $y_i = x_i'\beta + u_i$ 。
- 但在许多问题中，同一个个体或单位会同时对应多个方程。
- 典型例子包括似不相关回归（SUR）和面板数据模型。
- 联合估计的潜在好处：提高有效性、施加参数共享或跨方程约束、进行跨方程假设检验。

主要例子一：似不相关回归

SUR模型

研究家庭的食物、医疗和教育支出占比。对于家庭 i ：

$$y_i^j = \beta_j + \beta_j^F p_F + \beta_j^M p_M + \beta_j^E p_E + \beta_j^I I + u_i^j = x_i' \beta^j + u_i^j, \quad j = F, M, E$$

令 $x_i = [1, p_F, p_M, p_E, I]'$ ，并将三个方程堆叠：

$$Y_i = \begin{bmatrix} y_i^F \\ y_i^M \\ y_i^E \end{bmatrix}, \quad X_i = \begin{bmatrix} x_i' & 0 & 0 \\ 0 & x_i' & 0 \\ 0 & 0 & x_i' \end{bmatrix}_{3 \times 15}, \quad \beta = \begin{bmatrix} \beta^F \\ \beta^M \\ \beta^E \end{bmatrix}_{15 \times 1}, \quad U_i = \begin{bmatrix} u_i^F \\ u_i^M \\ u_i^E \end{bmatrix}$$

于是 $Y_i = X_i \beta + U_i$ 。SUR 的关键特征是：不同方程的误差项可能相关。

主要例子二：面板数据

面板数据

如果对同一个体 i 在 $t = 1, \dots, T$ 期重复观测，则每期都有一个方程：

$$y_{it} = x'_{it}\beta + u_{it}$$

$$Y_i = \begin{bmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{bmatrix}_{T \times 1}, \quad X_i = \begin{bmatrix} x'_{i1} \\ \vdots \\ x'_{iT} \end{bmatrix}_{T \times K}, \quad U_i = \begin{bmatrix} u_{i1} \\ \vdots \\ u_{iT} \end{bmatrix}_{T \times 1}$$

同样有 $Y_i = X_i\beta + U_i$ 。这里假设各期的系数均为 β 。

系统方程的一般形式

- 两类例子都可以写成 $Y_i = X_i\beta + U_i$ 。
- Y_i, U_i 是 $J \times 1$ 向量, X_i 是 $J \times K$ 矩阵。
- 在 SUR 中, J 是同一个个体对应的支出方程数; 在面板数据中, $J = T$ 是同一个体的时间期数。
- 不同系统 i 之间通常假设相互独立; 同一个系统内部的误差项可以相关。

系统普通最小二乘 (SOLS)

如果 $\mathbb{E}(X_i'U_i) = 0$, 则:

$$\beta = [\mathbb{E}(X_i'X_i)]^{-1} [\mathbb{E}(X_i'Y_i)]$$

用样本矩代替总体矩, 得到系统普通最小二乘 (SOLS):

$$\hat{\beta}^{\text{SOLS}} = \left[\sum_{i=1}^N X_i'X_i \right]^{-1} \left[\sum_{i=1}^N X_i'Y_i \right]$$

- 在 SUR 的例子中, SOLS 的点估计等价于逐方程 OLS。
- 在面板数据的例子中, SOLS 就是混合最小二乘 (POLS)。

SOLS 的大样本性质

记 $A = \mathbb{E}(X_i' X_i)$, $B = \mathbb{E}(X_i' U_i U_i' X_i)$ 。在一定正则条件下:

$$\sqrt{N}(\hat{\beta}^{\text{SOLS}} - \beta) \overset{a}{\sim} \mathcal{N}(0, A^{-1} B A^{-1})$$

$$\mathbb{V}(\widehat{\hat{\beta}^{\text{SOLS}}}) = \left[\sum_{i=1}^N X_i' X_i \right]^{-1} \left[\sum_{i=1}^N X_i' \hat{U}_i \hat{U}_i' X_i \right] \left[\sum_{i=1}^N X_i' X_i \right]^{-1}$$

其中 $\hat{U}_i = Y_i - X_i \hat{\beta}$ 。该公式已经允许同一个系统内部的误差项相关。

为什么需要聚类标准误？

- 许多数据具有组群结构：个体—年份、学生—班级、企业—地区等。
- 同一组内的误差项可能共享不可观测因素或受到共同冲击。
- 异方差稳健标准误只使用每条观测自己的残差平方，忽略组内残差之间的交叉乘积。
- 因此，组内相关时，异方差稳健标准误通常仍然不够稳健。

聚类相关结构

将全部观测按聚类排序，若不同聚类之间不相关，而同一聚类内部可以任意相关，则：

$$\mathbb{C}(u_h, u_l | X) = 0, \quad g(h) \neq g(l)$$

$$\mathbb{V}(U|X) = \begin{bmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_G \end{bmatrix}$$

- 每个 Ω_g 可以不同，允许组内异方差和任意相关。
- 非对角块为零表示不同聚类之间不相关；这是聚类标准误背后的关键结构假设。

聚类标准误

将第 g 组堆叠为 $Y_g = X_g\beta + U_g$, 则:

$$\sqrt{G}(\hat{\beta} - \beta) \overset{a}{\sim} \mathcal{N}(0, A^{-1}BA^{-1})$$

$$\widehat{\mathbb{V}}(\hat{\beta}) = \left[\sum_{g=1}^G X_g' X_g \right]^{-1} \left[\sum_{g=1}^G X_g' \hat{U}_g \hat{U}_g' X_g \right] \left[\sum_{g=1}^G X_g' X_g \right]^{-1}$$

- 中心极限定理作用在 G 个组得分 $X_g' U_g$ 上。
- 因此关键条件是聚类数 $G \rightarrow \infty$, 而不仅仅是总观测数很大。

聚类标准误与异方差稳健标准误

聚类方差的中间项可以展开为：

$$\sum_{g=1}^G X_g' \hat{U}_g \hat{U}_g' X_g = \underbrace{\sum_{g=1}^G \sum_{h \in g} \hat{u}_h^2 x_h x_h'}_{\text{异方差稳健项}} + \underbrace{\sum_{g=1}^G \sum_{h \in g} \sum_{l \in g, l \neq h} \hat{u}_h \hat{u}_l x_h x_l'}_{\text{组内协方差项}}$$

- 第一项就是异方差稳健标准误的中间项。
- 聚类标准误额外加入同一组内任意两条观测的残差交叉乘积。
- 如果组内误差不相关，第二项在总体中消失，两者渐近一致。

聚类层级如何选择？

- 由误差相关结构和处理分配层级决定，而不是机械地选择“越高越好”。
- 如果处理在同一村庄内相同，通常至少聚类到村庄；如果误差在省份内相关，村庄层级可能不足。
- 更高层级会减少聚类数 G ，从而使正态近似变差，也可能带来过度保守的推断。
- 若聚类数很少，应考虑 wild cluster bootstrap 等少聚类方法。

双向聚类

如果误差同时沿两个非嵌套维度相关，例如企业 g 和年份 t ，可以使用双向聚类：

$$\hat{V}_{tw} = \hat{V}_g + \hat{V}_t - \hat{V}_{g \cap t}$$

- $\hat{V}_g$ 、 $\hat{V}_t$ 分别为两个单向聚类方差。
- 减去交叉单元的方差是为了避免重复计算同时共享两个维度的观测对。
- 两个维度都需要足够多的聚类数；Stata 中 `reghdfe` 可以指定多个聚类变量。

抽样设计与聚类标准误

- 整群抽样：先抽取组，再调查组内个体；同组观测往往自然相关。
- 分层抽样：将总体分层，并在每一层内抽样；分层本身不自动意味着同层误差相关。
- 多阶段抽样：先抽取高层级单位，再在抽中的单位内继续抽样，通常形成层级相关结构。
- 抽样权重、分层变量和聚类变量是不同概念；聚类层级最终应由误差相关结构和研究设计决定。

Stata 示例：聚类到地区

下面的代码使用 `reghdfe`，并将标准误聚类到地区层面：

代码 1：地区层面的聚类标准误

```
1 // fixed_effects_ntv.do
2 clear
3 set more off
4 use datasets/NTV_Aggregate_Data_resshaped.dta
5 xtset tik_id year
6 forvalues i=1/5{
7     gen log_pop_`i'=logpop^`i'
8     gen log_wage_`i'=log_wage^`i'
9 }
10 // 显示结果
11 reghdfe Votes_Edinstvo_ Watch_probit if year==1999, absorb(region)
12     cluster(region)
13 // 显示控制变量
14 local E_controls "L.log_pop_* L.log_wage_* L.nurses L.doctors_pc"
15 reghdfe Votes_Edinstvo_ Watch_probit `E_controls' if year==1999,
16     absorb(region) cluster(region)
17 // 显示结果1995
```

GLS 的动机

- SOLS 的渐近方差包含 $\mathbb{E}(X_i' U_i U_i' X_i)$ 。
- 如果同一个系统内不同方程的误差项相关，SOLS 通常不是最有效的估计量。
- 记 $\Omega = \mathbb{E}(U_i U_i')$ ，它刻画方程之间或同一组内误差项的相关结构。
- 如果该结构已知或可以一致估计，可以利用它提高有效性。

广义最小二乘估计量

用 $\Omega^{-1/2}$ 左乘系统方程：

$$\Omega^{-1/2}Y_i = \Omega^{-1/2}X_i\beta + \Omega^{-1/2}U_i$$

令 $Y_i^* = \Omega^{-1/2}Y_i$, $X_i^* = \Omega^{-1/2}X_i$, 则 $\mathbb{V}(U_i^*) = I$ 。

$$\hat{\beta}^{\text{GLS}} = \left[\sum_{i=1}^N X_i' \Omega^{-1} X_i \right]^{-1} \left[\sum_{i=1}^N X_i' \Omega^{-1} Y_i \right]$$

GLS 的外生性代价

- GLS 的一致性要求 $\mathbb{E}(X_i' \Omega^{-1} U_i) = 0$ 。
- 当 Ω^{-1} 不是对角矩阵时，该条件会混合不同方程的解释变量和误差项。
- 因此，GLS 通常需要比 SOLS/POLS 更强的外生性假设：某一方程的误差项需要与所有方程的解释变量不相关。
- 如果只知道当期或本方程的外生性，POLS 往往更稳健。

GLS 的有效性

若 $\mathbb{E}(U|X) = 0$, $\mathbb{V}(U|X) = \bar{\Omega}$, 且相应矩阵可逆, 则给定 X 时, GLS 在所有线性无偏估计量中方差最小。

证明要点: 任取线性无偏估计量 $\tilde{\beta} = CY$, 无偏性意味着 $CX = I$ 。令 $A = (X'\bar{\Omega}^{-1}X)^{-1}X'\bar{\Omega}^{-1}$, $D = C - A$, 则 $DX = 0$, 从而

$$\mathbb{V}(\tilde{\beta}|X) - \mathbb{V}(\hat{\beta}^{\text{GLS}}|X) = D\bar{\Omega}D' \succeq 0$$

GLS 与 POLS: 效率与稳健性

在协方差结构正确时, GLS 与 POLS 的渐近方差满足:

$$[\mathbb{E}(X_i' \Omega^{-1} X_i)]^{-1} \preceq [\mathbb{E}(X_i' X_i)]^{-1} \mathbb{E}(X_i' \Omega X_i) [\mathbb{E}(X_i' X_i)]^{-1}$$

- GLS: 在 Ω 正确且外生性足够强时更有效。
- POLS: 不依赖 Ω 的设定, 对外生性要求较弱, 因而更稳健。
- 无论使用哪种点估计, 组内相关时推断都应使用聚类或其他稳健标准误。
- 若 SUR 各方程解释变量完全相同且没有跨方程约束, GLS 与逐方程 OLS 等价。

可行的广义最小二乘 (FGLS)

- 现实中 Ω 未知，需要先估计。
- 第一步：用 SOLS/POLS 得到一致估计和残差 $\hat{U}_i$ 。
- 第二步：计算 $\hat{\Omega} = \frac{1}{N} \sum_{i=1}^N \hat{U}_i \hat{U}_i'$ 。
- 第三步：用 $\hat{\Omega}$ 替代 Ω 计算 FGLS。
- 若 J 固定且 $\hat{\Omega} = \Omega + o_p(1)$ ，FGLS 与不可行 GLS 渐近等价。

一般线性约束与 Wald 检验

许多跨方程约束可以写成：

$$H_0 : R\beta = r, \quad \text{rank}(R) = q$$

若 $\sqrt{N}(\hat{\beta} - \beta) \overset{a}{\sim} \mathcal{N}(0, V)$ ，并且 $\hat{V}$ 是估计量的方差估计，则 Wald 统计量为：

$$W = (R\hat{\beta} - r)' (R\hat{V}R')^{-1} (R\hat{\beta} - r) \xrightarrow{D} \chi^2(q)$$

关键是 $\hat{V}$ 必须包含不同方程系数之间的协方差。

为什么必须联合估计？

- 例如在 SUR 中检验 $H_0 : \beta_E^F = \beta_F^E$ ，需要知道两个系数之差的方差。
- 该方差包含两个系数估计量之间的协方差。
- 如果逐方程分别估计，点估计可以得到，但跨方程协方差无法自然获得。
- 系统估计将各方程放在同一个估计框架中，从而可以直接构造 Wald 检验。

实例：OHIE 实验

OHIE 随机化平衡性检验

Finkelstein 等人（2012）在分析前检验处理组和对照组的随机化前特征是否平衡：

$$y_{ij} = \beta_{0j} + \beta_j \times \text{treatment}_{ij} + u_{ij}, \quad H_0 : \beta_1 = \cdots = \beta_J = 0$$

- 每个 j 对应一个随机化前特征。
- 原假设要求所有特征上的处理组—对照组差异同时为零。

OHIE 代码的结构逻辑

- 先将多个基线变量重命名为 outcome1、outcome2、.....。
- 再 reshape long: 每条记录对应一个家庭的一个特征变量，varid 标记变量编号。
- 回归中使用 i.varid 控制各特征的截距，使用 i.treatment#i.varid 生成每个特征的组间差异。
- 用 test 联合检验所有 treatment#varid 系数，并使用 cluster(household_id) 计算家庭层面聚类标准误。

OHIE 代码：整理为长格式

代码 2: OHIE: 整理为长格式

```
1 // ohie_qje_joint_test.do
2 clear frames
3 set more off
4 use datasets/OHIE_QJE.dta
5
6 // 整理为长格式
7 local lottery_list "birthyear_list female_list english_list self_list
   first_day_list have_phone_list pobox_list zip_msa"
8 // 整理为长格式
9 frames put `lottery_list' treatment weight_12m household_id numhh_list, into(
   joint_compare)
10 frame change joint_compare
11 local i=0
12 foreach v of varlist `lottery_list'{
13     local i=`i'+1
14     rename `v' outcome`i'
15 }
16 gen id=_n
17 reshape long outcome, i(id) j(varid)
18 local test_coefs ""
```

OHIE 代码：回归与联合检验

代码 3: OHIE: 回归与联合检验

```
1  #####
2  forvalues j=1/`i'{
3      local test_coefs "`test_coefs' 1.treatment#`j'.varid"
4  }
5  di "`test_coefs'"
6  reg outcome i.varid i.treatment#i.varid i.numhh_list#i.varid , cl(household_id)
7  test `test_coefs'
8  frame change default
```

为什么 reshape 后手动检验，而不是直接 sureg?

- 该例的目标是联合平衡性检验，而不是通过 SUR/FGLS 提高各方程系数的效率。
- reshape 后，每个待检验系数都直接对应一个 treatment#varid 交互项，检验形式直观。
- 堆叠回归配合 cluster(household_id)，正好对应同一家庭多条记录相关的推断结构。
- sureg 也能处理多方程和跨方程检验；这里的区别是模型结构、协方差估计和检验系数构造方式，而不是 sureg 原则上不可行。