第10章 面板数据

面板数据的特点

- 面板数据(panel data 或 longitudinal data)，指在一段时间内跟踪同一组个体(individual)的数据。
- 它既有横截面维度(n位个体)，又有时间维度(T 个时期)。
- 一个T=3的面板数据结构如表

	y	x_1	x_2	x_3
个体 1: $t=1$				
个体 1: $t=2$				
个体 1: $t=3$				
个体 2: $t=1$				
个体 2: $t=2$				
个体 2: $t=3$				
……				
个体n: $t=1$				
个体n: $t=2$				
个体n: $t=3$				

- 通常的面板数据T 较小，n较大，在使用大样本理论时让n趋于无穷大，称为“短面板”(short panel)。
- 如果T 较大，n较小，则称为“长面板”(long panel)。
- 在面板模型中，如果解释变量包含被解释变量的滞后值，称为“动态面板”(dynamic panel)；
- 反之，称为“静态面板”(static panel)。
- 在面板数据中，如果每个时期在样本中的个体完全一样，则称为“平衡面板”(balanced panel)；
- 反之，则称为“非平衡面板”(unbalanced panel)。
- 主要关注平衡面板。

- 面板数据的主要优点：
- (1) 有助于解决遗漏变量问题
- 遗漏变量常由不可观测的个体差异或“异质性”(heterogeneity)造成(比如个体能力)。
- 如果个体差异“不随时间而改变”(time invariant)，则面板数据提供了解决遗漏变量问题的又一利器。
- (2) 提供更多个体动态行为的信息
- 面板数据有横截面与时间两个维度，可解决截面数据或时间序列不能解决的问题。
- (3) 样本容量较大
- 同时有截面与时间维度，面板数据的样本容量通常更大，可提高估计精度。

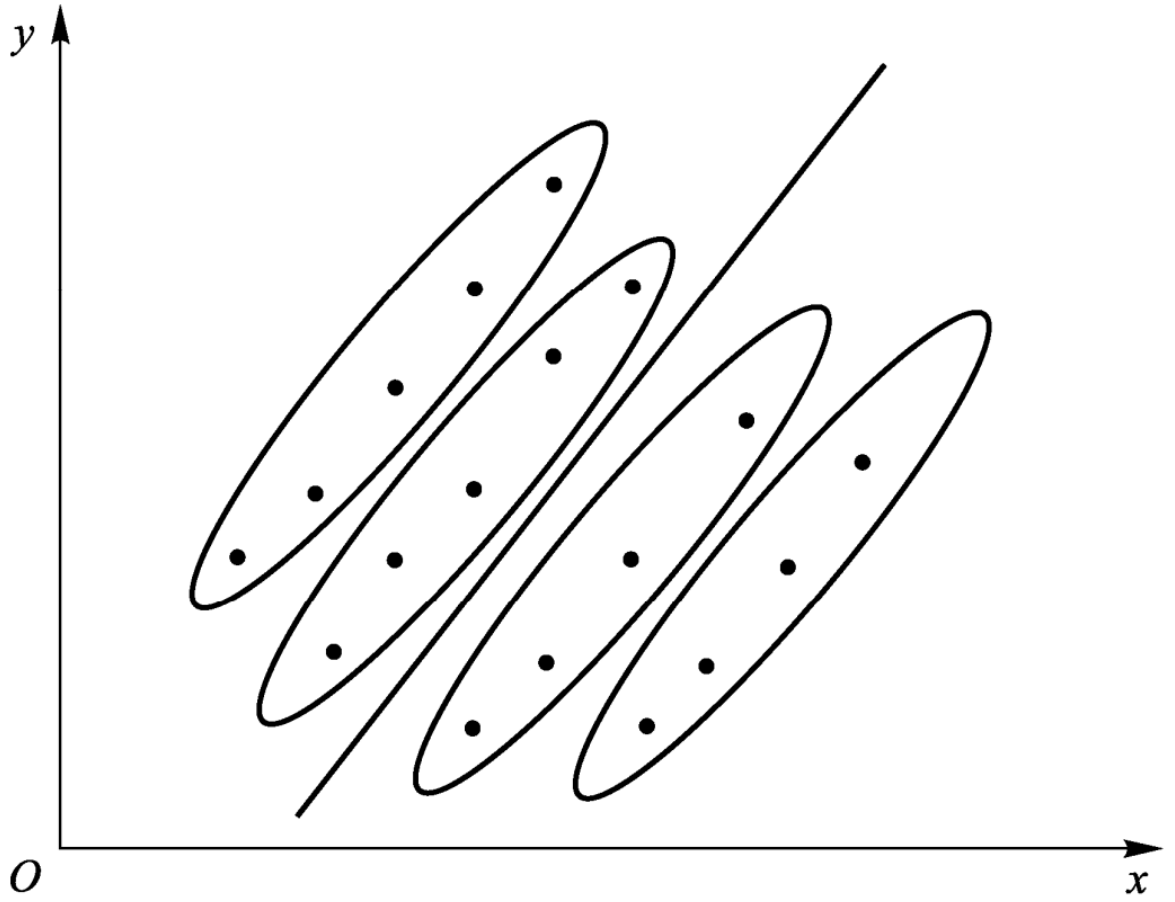
- **例** 如何区分规模效应与技术进步对企业生产效率的影响。截面数据没有时间维度，无法观测到技术进步。单个企业的时间序列数据，也无法区分生产效率提高有多少由于规模扩大，有多少由于技术进步。
- **例** 对于失业问题，截面数据告诉我们在某个时点上哪些人失业，时间序列告诉我们某个人就业与失业的历史，但均无法告诉我们是否失业的总是同一批人(低流转率)，还是失业的人群总在变动(高流转率)。
- 如有面板数据，就可能解决上述问题。

- 面板数据也会带来问题。
- 样本数据通常不满足 iid 假定，因为同一个体在不同期的扰动项一般存在自相关。
- 面板数据的收集成本通常较高，不易获得。

面板数据的估计策略

- 一个极端策略是，将面板看成截面数据进行混合回归(pooled regression)，即要求样本中每位个体拥有完全相同的回归方程。
- 混合回归的缺点是，忽略个体不可观测的异质性(heterogeneity)，而该异质性可能与解释变量相关，导致估计不一致。
- 另一极端策略是，为每位个体估计单独的回归方程。
- 分别回归的缺点是，忽略个体的共性，可能没有足够大的样本容量。
- 实践中常采用折衷的策略，即假定个体的回归方程拥有相同的斜率，但可有不同截距项，以捕捉异质性

- 面板数据中不同个体的截距项可以不同



- 这种模型称为“个体效应模型”(individual-specific effects model)：

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it} \quad (i = 1, \cdots, n; t = 1, \cdots, T)$$

- $\boldsymbol{z}_i$为不随时间而变(time invariant)的个体特征（$\boldsymbol{z}_{it} = \boldsymbol{z}_i, \forall t$），比如性别；
- $\boldsymbol{x}_{it}$可以随个体及时间而变(time-varying)。
- 扰动项由$(u_i + \varepsilon_{it})$两部分构成，称为“复合扰动项”(composite
- error term)。
- 不可观测的随机变量u_i是代表个体异质性的截距项，即“个体效应”(individual effects)。
- ε_{it}为随个体与时间而改变的扰动项，称为“idiosyncratic error”
- 一般假设$\{\varepsilon_{it}\}$为独立同分布，且与u_i不相关。

- 如果u_i与某个解释变量相关，则进一步称为“**固定效应模型**”(Fixed Effects Model，简记 FE)。
- 此时 OLS 不一致。
- 解决方法是转换模型，消去u_i获得一致估计。
- 如果u_i与所有解释变量$(\boldsymbol{x}_{it}, \boldsymbol{z}_i)$均不相关，则进一步称为“**随机效应模型**”(Random Effects Model，简记 RE)。
- 与横截面数据相比，面板数据提供了更丰富的模型与估计方法。

混合回归

- 对于个体效应模型

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{i}}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it}$$

- 如果所有个体都拥有完全一样的回归方程，则$u_1 = u_2 = \cdots u_n$。
- 将相同的个体效应统一记为α，方程可写为

$$y_{it} = \alpha + \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{\boldsymbol{i}}^{'}\boldsymbol{\delta} + \varepsilon_{it}$$

  其中，$\boldsymbol{x}_{it}$不包括常数项。

- 把所有数据放在一起，像横截面数据那样进行 OLS 回归，故称“**混合回归**”(pooled regression)。

- 虽可假设不同个体的扰动项相互独立，但同一个体在不同时期的扰动项之间往往自相关。
- 每位个体不同时期的所有观测值构成一个“**聚类**”(cluster)。
- 样本观测值可分为不同的聚类，在同一聚类里的观测值互相相关，不同聚类之间的观测值不相关，称为“**聚类样本**”(cluster sample)。
- 对于聚类样本，仍可进行 OLS 估计，但需使用“聚类稳健的标准误”(cluster-robust standard errors)，形式上也是夹心估计量，表达式更为复杂。
- 对于样本容量为nT的平衡面板，共有n个聚类，而每个聚类中包含T期观测值。

- 使用聚类稳健标准误的前提是，聚类中的观测值数目T较小，而聚类数目n较大$(n \to \infty)$。
- 此时聚类稳健标准误是真实标准误的一致估计。
- 聚类稳健标准误更适用于时间维度 T 比截面维度n小的**短面板**。
- 在推导过程中未假定同方差，故聚类稳健标准误也是异方差稳健的。
- 混合回归的基本假设是不存在个体效应，对此须进行统计检验。

到底cluster在哪一层级？

- 1. 面板数据，至少应控制在个体层面。
- 2. x和y不在同一level：例如省管县对企业绩效的影响
- 3. 横截面数据：一般是cluster在一个更高的范围内。多高呢？Tradeoff
  - cluster的数目太少，怎么办？ 一般应大于42
  - 可以多报几种cluster的方式，表明稳健性

When number of clusters is small

- Imbens G., KolesárM.2016. Robust Standard Errors in Small Samples: Some Practical Advice. *Review of Economics and Statistics*, 98(4):701-712.
- Wild bootstrap standard errors
- Statacode: cgmwildboot

固定效应模型：组内估计量

- 考虑固定效应模型：

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it}$$

其中，u_i与某解释变量相关，故OLS不一致。

- 解决方法：通过模型变换，消掉个体效应u_i。

- 给定个体i，方程两边对时间取平均：

$$\bar{y}_i = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i$$

其中，$\bar{y}_i \equiv \frac{1}{T}\sum_{t=1}^{T} y_{it}$，$\bar{\boldsymbol{x}}_i$与$\bar{\varepsilon}_i$的定义类似。

- 将原方程减去平均方程，可得离差形式：

$$y_{it} - \bar{y}_i = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i)$$

- $\boldsymbol{z}_i$与u_i被消去。
- 定义$\tilde{y}_{it} \equiv y_{it} - \bar{y}_i$，$\tilde{\boldsymbol{x}}_{it} \equiv \boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i$，$\tilde{\varepsilon}_{it} \equiv \varepsilon_{it} - \bar{\varepsilon}_i$，则

$$\tilde{y}_{it} = \tilde{\boldsymbol{x}}_{it}'\boldsymbol{\beta} + \tilde{\varepsilon}_{it}$$

- 只要新扰动项$\tilde{\varepsilon}_{it}$与新解释变量$\tilde{\boldsymbol{x}}_{it}$不相关，则OLS一致，称为“固定效应估计量”(Fixed Effects Estimator)，记为$\hat{\boldsymbol{\beta}}_{FE}$。
- $\hat{\boldsymbol{\beta}}_{FE}$主要使用每位个体的组内离差信息，也称“组内估计量”(within estimator)。

- 即使u_i与$\boldsymbol{x}_{it}$相关，只要使用组内估计量，即可得一致估计，这是面板数据的一大优势。
- 由于可能存在组内自相关，应使用以每位个体为聚类的聚类稳健标准误。
- 在离差变换过程中，$\boldsymbol{z}_i'\boldsymbol{\delta}$也消掉，无法估计$\boldsymbol{\delta}$。
- $\widehat{\boldsymbol{\beta}}_{FE}$无法估计不随时间而变的变量之影响，这是FE的一大缺点。
- 为保证$(\varepsilon_{it}-\bar{\varepsilon}_i)$与$(\boldsymbol{x}_{it}-\bar{\boldsymbol{x}}_i)$不相关，须假定个体$i$满足严格外生性（比前定变量或同期外生的假定更强），即$E(\varepsilon_{it}|\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iT})=0$，因为$\bar{\boldsymbol{x}}_i$中包含了所有$(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{iT})$的信息。
- 严格外生性的假定，比前定变量或同期外生的假定更强。

固定效应模型：LSDV 法

- 考虑固定效应模型：

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it}$$

- 个体固定效应u_i，传统上视为个体i的待估参数，即个体i的截距项。
- 对于n位个体的n个不同截距项，可在方程中引入$(n-1)$个个体虚拟变量来体现：

$$y_{it} = \alpha + \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + \sum_{i=2}^{n}\gamma_i D_i + \varepsilon_{it}$$

- 其中，个体虚拟变量$D_2 = 1$，如果为个体2；否则，$D_2 = 0$。其他$(D_3, \cdots, D_n)$的定义类似。
- 用OLS 估计此方程，称为“最小二乘虚拟变量法”(Least Square Dummy Variable，LSDV)。

- LSDV 法的估计结果与组内估计量 FE 完全相同。
- 正如线性回归与离差形式的回归在某种意义上等价

$$y_i = \alpha + \beta x_i + \varepsilon_i \;\Leftrightarrow\; y_i - \bar{y} = \beta(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$$

- 做完LSDV后，如发现某些个体的虚拟变量不显著而删去，则LSDV的结果就不会与 FE 相同。
- LSDV 的好处是，可得到对个体异质性u_i的估计。
- LSDV 法的缺点是，如果n很大，须在回归方程中引入很多虚拟
- 变量，可能超出 Stata 所允许的变量个数。

固定效应模型：一阶差分法

- 考虑固定效应模型：

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it}$$

- 其一阶滞后为：

$$y_{i,t-1} = \boldsymbol{x}_{i,t-1}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{i,t-1}$$

- 对于固定效应模型，还可对原方程两边进行一阶差分，消去个体效应u_i：

$$y_{it} - y_{i,t-1} = (\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})^{'}\boldsymbol{\beta} + (\varepsilon_{it} - \varepsilon_{i,t-1})$$

- 使用 OLS 即得到“一阶差分估计量”(First Differencing Estimator)，记为$\widehat{\boldsymbol{\beta}}_{FD}$。
- 只要扰动项的一阶差分$(\varepsilon_{it} - \varepsilon_{i,t-1})$与解释变量的一阶差分$(\boldsymbol{x}_{it} - \boldsymbol{x}_{i,t-1})$不相关，则$\widehat{\boldsymbol{\beta}}_{FD}$一致。
- 此一致性条件比保证$\widehat{\boldsymbol{\beta}}_{FE}$一致的严格外生性假定更弱。

- 可以证明，如果$T = 2$，则$\hat{\boldsymbol{\beta}}_{FD} = \hat{\boldsymbol{\beta}}_{FE}$。
- 对于$T > 2$，如果$\{\varepsilon_{it}\}$为独立同分布，则$\hat{\boldsymbol{\beta}}_{FE}$比$\hat{\boldsymbol{\beta}}_{FD}$更有效率。
- 实践中，主要用$\hat{\boldsymbol{\beta}}_{FE}$，较少用$\hat{\boldsymbol{\beta}}_{FD}$。

时间固定效应

- 个体固定效应模型解决了不随时间而变(time invariant)但随个体而异的遗漏变量问题。
- 还可能存在不随个体而变(individual invariant)，但随时间而变(time varying)的遗漏变量问题。
- 比如，企业经营的宏观经济环境。
- 在个体固定效应模型中加入时间固定效应(λ_t):

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + \lambda_t + u_i + \varepsilon_{it}$$

其中，λ_t随时间而变，但不随个体而变。

- 可视λ_t为第t期特有的截距项，并解释为“第t期”对y的效应；故称$\{\lambda_1, \cdots, \lambda_T\}$为“时间固定效应”（time fixed effects）。

- 使用 LSDV 法，对每个时期定义一个虚拟变量，把$(T-1)$个时间虚拟变量包括在回归方程中：

$$y_{it} = \alpha + \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + \sum_{t=2}^{T} \gamma_t D_t + u_i + \varepsilon_{it}$$

- 时间虚拟变量，$D_2 = 1$，如果为$t = 2$；否则，$D_2 = 0$；以此类推。
- 该方程既考虑了个体固定效应，又考虑了时间固定效应，称为“双向固定效应”(Two-way FE)。
- 可通过检验这些时间虚拟变量的联合显著性来判断是否应使用双向固定效应模型。
- 如果仅考虑个体固定效应，称为“单向固定效应”(One-way FE)。

- 有时为节省参数(比如，时间维度T 较大)，可引入时间趋势项，以替代上述$(T-1)$ 个时间虚拟变量：

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + \gamma t + u_i + \varepsilon_{it}$$

其中，时间趋势项$t = 1,2,3,4,5,\cdots$

- 上式隐含假定，每个时期的时间效应相等，即每期均增加γ。
- 如果此假定不太可能成立，应在方程中加入时间虚拟变量。

随机效应模型

- 考虑随机效应模型：

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it}$$

- 其中，个体效应u_i与解释变量均不相关，故 OLS 一致。
- 由于扰动项由$(u_i + \varepsilon_{it})$组成，不是球型扰动项，故 OLS 不是最有效率的。
- 希望进行更有效率的FGLS估计。

- 假设不同个体之间的扰动项互不相关。
- 由于u_i的存在，同一个体不同时期的扰动项之间仍存在自相关。
- 对于$t \neq s$，可证明

$$\mathrm{Cov}(u_i + \varepsilon_{it}, u_i + \varepsilon_{is}) = \mathrm{Cov}(u_i, u_i) + \underbrace{\mathrm{Cov}(u_i, \varepsilon_{is})}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_{it}, u_i)}_{=0} + \underbrace{\mathrm{Cov}(\varepsilon_{it}, \varepsilon_{is})}_{0}$$

$$= \mathrm{Var}(u_i) \equiv \sigma_u^2 \neq 0$$

- 其中，$\sigma_u^2 \equiv Var(u_i)$为个体效应$u_i$的方差。

- 如果$t = s$，则

$$Var(u_i + \varepsilon_{it}) = \sigma_u^2 + \sigma_\varepsilon^2$$

- 其中，$\sigma_\varepsilon^2 \equiv Var(\varepsilon_{it})$为$\varepsilon_{it}$的方差（不随$i$，$t$变化）。
- 当$t \neq s$时，个体i扰动项的自相关系数为

$$\rho \equiv Corr(u_i + \varepsilon_{it}, u_i + \varepsilon_{is}) \equiv \frac{Cov(u_i + \varepsilon_{it}, u_i + \varepsilon_{is})}{Var(u_i + \varepsilon_{it})} = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2}$$

- 自相关系数ρ越大，则复合扰动项$(u_i + \varepsilon_{it})$中个体效应的部分（u_i）越重要。
- 由于扰动项$(u_i + \varepsilon_{it})$存在组内自相关，故 OLS 不是最有效率的。

- 目标：转换模型，使变换后扰动项无自相关，进行GLS估计
- 定义

$$\theta \equiv 1 - \frac{\sigma_\varepsilon}{(T\sigma_u^2 + \sigma_\varepsilon^2)^{\frac{1}{2}}} = 1 - \left(\frac{\sigma_\varepsilon^2}{T\sigma_u^2 + \sigma_\varepsilon^2}\right)^{1/2}$$

- 其中，T为面板数据的时间维度。
- 显然，$0 \leq \theta \leq 1$。

$$y_{it} = \boldsymbol{x}_{it}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \varepsilon_{it}$$

- 给定个体i，将方程两边对时间进行平均：

$$\bar{y}_i = \bar{\boldsymbol{x}}_{i}^{'}\boldsymbol{\beta} + \boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i$$

- 然后同乘θ：

$$\theta\bar{y}_i = \theta\bar{\boldsymbol{x}}_{i}^{'}\boldsymbol{\beta} + \theta\boldsymbol{z}_{i}^{'}\boldsymbol{\delta} + \theta u_i + \theta\bar{\varepsilon}_i$$

- 两方程相减可得“广义离差”(quasi-demeaned)模型：

$$y_{it} - \theta\bar{y}_i = (x_{it} - \theta\bar{x}_i)'\boldsymbol{\beta} + (1-\theta)\boldsymbol{z}_i'\boldsymbol{\delta} + \underbrace{[(1-\theta)u_i + (\varepsilon_{it} - \theta\bar{\varepsilon}_i)]}_{\text{扰动项}}$$

- 由于$0 \leq \theta \leq 1$，故$(y_{it} - \theta\bar{y}_i)$只是减去$\bar{y}_i$的一部分，故名“广义离差”。
- 广义离差方程的扰动项$[(1-\theta)u_i + (\varepsilon_{it} - \theta\bar{\varepsilon}_i)]$不再有自相关。
- 对此方程进行 OLS 估计，即为 GLS 估计量。
- 但θ通常未知，须先估计$\hat{\theta}$，再进行FGLS估计。

- 可用下式来估计$\hat{\theta}$

$$\hat{\theta} \equiv 1 - \frac{\hat{\sigma}_{\varepsilon}}{(T\hat{\sigma}_{u}^{2} + \hat{\sigma}_{\varepsilon}^{2})^{1/2}}$$

- 其中，$\hat{\sigma}_{u}$与$\hat{\sigma}_{\varepsilon}$分别为$\sigma_{u}$和$\sigma_{\varepsilon}$的样本估计值。
- 对于随机效应模型，由于 OLS 一致，且其扰动项为$(u_i + \varepsilon_{it})$，故可用OLS的残差估计$(\sigma_u^2+\sigma_\varepsilon^2)$。
- 另一方面，FE 也一致，且其扰动项为$(\varepsilon_{it} - \bar{\varepsilon}_i)$，故可用FE的残差估计$\sigma_\varepsilon^2$。
- 由此得到$\hat{\theta}$，再用FGLS估计原模型，可得“随机效应估计量”(Random Effects Estimator)，记为$\hat{\boldsymbol{\beta}}_{RE}$。
- 如假设扰动项服从正态分布，可进行最大似然估计(MLE)。

组间估计量

- 对于随机效应模型，还可使用“组间估计量”。
- 如每位个体的时间序列数据较不准确或噪音较大，可对每位个体取时间平均值，然后用平均值作**横截面回归**：

$$\bar{y}_i = \bar{\boldsymbol{x}}_i^{'}\boldsymbol{\beta} + \boldsymbol{z}_i^{'}\boldsymbol{\delta} + u_i + \bar{\varepsilon}_i \ \ (i = 1, \cdots, n)$$

- 对上式用 OLS，即为“组间估计量”(Between Estimator)，记 $\widehat{\boldsymbol{\beta}}_{BE}$ 。
- 由于 $\{\bar{\boldsymbol{x}}_i, \boldsymbol{z}_i\}$包含$\{\boldsymbol{x}_i, \boldsymbol{z}_i\}$的信息，如$u_i$与$\{\boldsymbol{x}_{it}, \boldsymbol{z}_i\}$相关，则$\widehat{\boldsymbol{\beta}}_{BE}$不一致。
- 故不能在固定效应模型下使用组间估计法。
- 即使在随机效应模型下，由于面板数据被压缩为截面数据，损失较多信息量，组间估计法也不常用。

拟合优度的度量

- 对于面板模型，如使用混合回归，可直接用混合回归的R^2衡量拟合优度。
- 如使用固定效应或随机效应，拟合优度的度量略复杂。
- 对于有常数项的线性回归模型，拟合优度R^2等于被解释变量y与预测值$\hat{y}$之间相关系数的平方，即$R^2 = [corr(y, \hat{y})]^2$。

- 给定估计量$(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}})$，Stata 提供了以下三种$R^2$。
- (1)对应于原模型$y_{it}=\boldsymbol{x}_{it}'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+u_i+\varepsilon_{it}$，称$\left[corr(y_{it},\boldsymbol{x}_{it}'\widehat{\boldsymbol{\beta}}+\boldsymbol{z}_i'\widehat{\boldsymbol{\delta}})\right]^2$为“整体$R^2$”（$R^2$ overall），衡量估计量$(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}})$对原模型的拟合优度。
- (2)对应于组内模型$\tilde{y}_{it}=\tilde{\boldsymbol{x}}_{it}'\boldsymbol{\beta}+\tilde{\varepsilon}_{it}$，称$\left[corr(\tilde{y}_{it},\tilde{\boldsymbol{x}}_{it}'\widehat{\boldsymbol{\beta}})\right]^2$为“组内$R^2$（$R^2$ within），衡量估计量$(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}})$对组内模型的拟合优度。
- (3)对应于组间模型$\bar{y}_i=\bar{\boldsymbol{x}}_i'\boldsymbol{\beta}+\boldsymbol{z}_i'\boldsymbol{\delta}+u_i+\bar{\varepsilon}_i$，称$\left[corr(\bar{y}_i,\bar{\boldsymbol{x}}_i'\widehat{\boldsymbol{\beta}}+\boldsymbol{z}_i'\widehat{\boldsymbol{\delta}})\right]^2$为“组间$R^2$（$R^2$ between），衡量估计量$(\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\delta}})$对组间模型的拟合优度。

- 无论固定效应、随机效应还是组间回归，都可计算这三种R^2。
- 对于固定效应模型，建议使用组内R^2。
- 对于组间回归模型，建议使用组间R^2。
- 对于随机效应模型，这三种R^2都只是相应的相关系数平方而已(并非随机效应模型广义离差回归的R^2)。

非平衡面板

- 在面板数据中，如每个时期在样本中的个体完全一样，称为“平衡面板数据” (balanced panel)。
- 但有时某些个体的数据可能缺失(比如，个体死亡、企业倒闭或被兼并、个体不再参与调查)，或者新个体在后来才加入到调查中来。
- 如每个时期观测到的个体不完全相同，称为“非平衡面板”(unbalanced panel)或“不完全面板”(incomplete panel)。
- 非平衡面板数据不影响计算离差形式的组内估计量(within estimator)，固定效应模型的估计可照样进行。
- 对于随机效应模型，非平衡面板数据也无实质影响。

- 假设个体i的时间维度为T_i，只要在做广义离差变换时，为每位个体定义

$$\hat{\theta}_i \equiv 1 - \frac{\hat{\sigma}_\varepsilon}{(T_i\hat{\sigma}_u^2 + \hat{\sigma}_\varepsilon^2)^{1/2}}$$

- 即可照常进行 FGLS 估计。
- 非平衡面板可能出现的最大问题是，那些原来在样本中但后来丢掉的个体，如果“丢掉”的原因内生(与扰动项相关)，则会导致样本不具有代表性(不再是随机样本)，导致估计量不一致。
- 比如，低收入人群更易从面板数据中丢掉。
- 如果从非平衡面板数据中提取一个平衡的面板数据子集，则必然会损失样本容量，降低估计效率。
- 如人为“丢掉”个体并非完全随机，同样会破坏样本的随机性。

究竟该用固定效应还是随机效应模型

- 处理面板数据，究竟使用固定效应还是随机效应是根本问题。
- 检验原假设“H_0: u_i与 $\boldsymbol{x}_{it}, \boldsymbol{z}_i$不相关”(随机效应为正确模型)。
- 如果H_0成立，则FE和RE都已知，但 RE 比 FE 更有效率。
- 如果H_0不成立，则FE 一致，而 RE 不一致。
- 如果H_0成立，则 FE 与 RE 估计量将共同收敛于真实的参数值，二者的差距将在大样本下消失，故$(\widehat{\boldsymbol{\beta}}_{FE} - \widehat{\boldsymbol{\beta}}_{RE}) \xrightarrow{p} \mathbf{0}$。
- 反之，如果二者的差距过大，则倾向于拒绝原假设。

- 以二次型度量此距离，豪斯曼检验(Hausman, 1978)的统计量为

$$(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\hat{\boldsymbol{\beta}}_{\mathrm{RE}})'\left[\widehat{\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\hat{\boldsymbol{\beta}}_{\mathrm{RE}})}\right]^{-1}(\hat{\boldsymbol{\beta}}_{\mathrm{FE}}-\hat{\boldsymbol{\beta}}_{\mathrm{RE}})\xrightarrow{d}\chi^2(K)$$

- 其中，K为$\widehat{\boldsymbol{\beta}}_{FE}$的维度，即 $\boldsymbol{x}_{it}$中随时间而变的解释变量个数。
- 如果该统计量大于临界值，则拒绝H_0。
- 此检验的缺点是，为计算$\widehat{Var(\widehat{\boldsymbol{\beta}}_{FE}-\widehat{\boldsymbol{\beta}}_{RE})}$，假设在$H_0$成立情况下，$\widehat{\boldsymbol{\beta}}_{RE}$是最有效率的(fully efficient)。
- 但如果扰动项存在异方差，则$\widehat{\boldsymbol{\beta}}_{RE}$并非最有效率。
- 传统的豪斯曼检验不适用于异方差的情形，须使用异方差稳健的豪斯曼检验。

- Stata命令
- xtreg

  xtset province year

固定效应

  xtreg y x1 x2 x3,fe r

双向固定效应

 xtreg y x1 x2 x3 i.year,fe r

- reghdfe

reghdfe depvar [indepvars] [if] [in] [weight], absorb(absvars) [options]

https://mp.weixin.qq.com/s/fKTSMyZLnhFCAUd_LaWnVQ