面板数据模型

李德山

[1] 西南科技大学 经济管理学院

[2]School of Economics and Management
Southwest University of Science and Technology

2021 年 3 月 19 日

主要内容

主要内容

1. 面板数据
2. 面板数据模型因果关系分析的直观理解

主要内容

1. 面板数据
2. 面板数据模型因果关系分析的直观理解
3. 固定效应模型估计方法

主要内容

1. 面板数据
2. 面板数据模型因果关系分析的直观理解
3. 固定效应模型估计方法
4. 动态面板的 GMM 估计方法

主要内容

1. 面板数据
2. 面板数据模型因果关系分析的直观理解
3. 固定效应模型估计方法
4. 动态面板的 GMM 估计方法
5. 面板数据模型常见问题

面板数据

- **面板数据：**包含多个个体，并且同一个体有一系列不同时间观测点的数据。
  - 个体可以是个人、企业、行业或国家；时间可以是日、月、季或年。

ID	YEAR	INC	EDU	AGE	GENDER
1	2017	800	3	23	1
1	2018	1000	4	24	1
1	2019	1200	5	25	1
2	2017	1200	5	30	0
2	2018	1250	6	31	0
2	2019	1300	7	32	0

Figure: 面板数据结构

面板数据

- 然而并不是所有包含个体和时间两个维度的数据都是面板数据。下面就是一个合并的横截面数据。

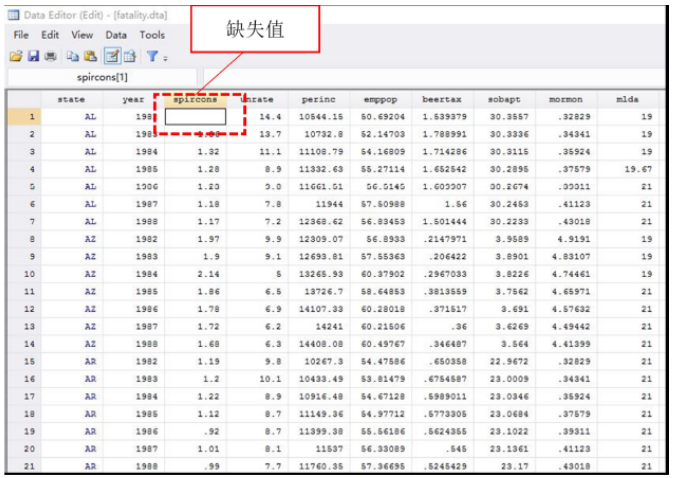
YEAR	INC	EDU	AGE	GENDER
2017	800	3	23	1
2018	1000	4	24	1
2019	1200	5	25	1
2017	1200	5	30	0
2018	1250	6	31	0
2019	1300	7	32	0

Figure: 合并横截面数据结构

- 它并没有跟踪记录同一个个体。只是多年横截面数据数据的简单叠加。

面板数据

- **面板数据分类：**平衡面板与非平衡面板
  - 在面板数据中，如果对于每个个体，它们都有相同时间 T 的观测点，称之为平衡面板数据。如果面板中至少有一个时间、一个个体的观测值是缺失的，则称为非平衡面板数据。



	state	year	spircons	unrate	perinc	emppop	beertax	sobapt	mormon	mlda
1	AL	1981	.	14.4	10544.15	50.69204	1.539379	30.3557	.32829	19
2	AL	1982	.	13.7	10732.8	52.14703	1.788991	30.3336	.34341	19
3	AL	1984	1.92	11.1	11108.79	54.16809	1.714286	30.3115	.35924	19
4	AL	1985	1.28	8.9	11332.63	55.27114	1.652542	30.2895	.37579	19.67
5	AL	1986	1.23	9.8	11661.51	56.5145	1.609907	30.2674	.39311	21
6	AL	1987	1.18	7.8	11944	57.50988	1.56	30.2453	.41123	21
7	AL	1988	1.17	7.2	12368.62	58.83453	1.501444	30.2233	.43018	21
8	AZ	1982	1.97	9.9	12309.07	56.8933	.2147971	3.9589	4.9191	19
9	AZ	1983	1.9	9.1	12693.81	57.55363	.206422	3.8901	4.83107	19
10	AZ	1984	2.14	8	13265.93	60.37902	.2967033	3.8226	4.74461	19
11	AZ	1985	1.86	6.5	13726.7	58.64853	.3813559	3.7562	4.65971	21
12	AZ	1986	1.78	6.9	14107.33	60.28018	.371617	3.691	4.57632	21
13	AZ	1987	1.72	6.2	14241	60.21506	.36	3.6269	4.49442	21
14	AZ	1988	1.68	6.3	14408.08	60.49767	.346487	3.564	4.41399	21
15	AR	1982	1.19	9.8	10267.3	54.47586	.650358	22.9672	.32829	21
16	AR	1983	1.2	10.1	10433.49	53.81479	.6754587	23.0009	.34341	21
17	AR	1984	1.22	8.9	10916.48	54.67128	.5989011	23.0346	.35924	21
18	AR	1985	1.12	8.7	11149.36	54.97712	.5773305	23.0684	.37579	21
19	AR	1986	.92	8.7	11399.38	55.56186	.5624366	23.1022	.39311	21
20	AR	1987	1.01	8.1	11537	56.33089	.545	23.1361	.41123	21
21	AR	1988	.99	7.7	11760.35	57.36695	.5245429	23.17	.43018	21

  - 如果非平衡面板数据缺失是由于非随机原因造成的，我们必须考虑缺失的原因。
  - 例如，收入跟踪调查数据。“懒惰”的人比较容易选择退出数据调查，因为懒惰会影响收入，造成样本选择偏差。

面板数据

- **面板数据分类：**短面板与长面板

- 短面板是指个体维度 N 较大，时间维度 T 较小。例如，大型微观调查数据可能跟踪几百万人，但每 5 年调查一次，20 年的数据 T 也只有 4。
- 长面板是指数据的 N 较小，T 较大。例如，G7 国家经济数据只包含 7 个国家，但是有上百年的数据。
- 有些数据 N 比较大，T 也较大，则称为“大面板”数据。

面板数据

- **面板数据分类：**静态面板与动态面板
  - 在面板数据模型中，如果解释变量包含被解释变量的滞后值，则称为动态面板。
  - 反之，称为静态面板数据。

面板数据

- **面板数据的主要优点**

1. 有助于解决遗漏变量问题。如果不可观测的个体差异“不随时间而改变”，则面板数据提供了解决这种类型的遗漏变量的一种办法。
2. 提供更多个体动态行为的信息。可解决截面数据或时间序列不能解决的一些问题。例如，研究技术进步对企业生产效率的影响，截面数据无法观测到技术进步。
3. 样本容量较大。可提供估计精度。

- 面板数据通常不满足 $i.i.d.$ 假定，因为同一个个体在不同期的扰动项一般存在自相关。

- **面板数据的信息来源。**

- 面板数据包含了两个维度的信息，其变量的变化来源包含了不同个体间的差异和同一个体在不同时间上的差异。总方差可以分解为个体间方差（组间方差）和个体内方差（组内方差）：

$$TotalVariation = BetweenVarition + WithinVarition$$

$$s_T^2 = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\bar{X})^2$$

$$s_B^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{X}_i-\bar{X})^2$$

$$s_W^2 = \frac{1}{NT-1}\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\bar{X}_i)^2$$

面板数据

- 其中，个体间方差是个体平均值的方差，个体内方差是相同个体在不同时间上值的方差。
- 由于是计算样本的方差，做了 NT-1 和 N-1 的调整。因此，样本总方差只是近似等于二者之和，$s_T^2 \approx s_B^2 + s_W^2$。

面板数据模型因果关系分析的直观理解

- **为了方便直观理解，假设要估计教育对收入的因果影响，但影响收入 INC 的因素除了教育 EDU，还有性别 GENDER，个人的先天能力 TALENT 和运气 LUCK。**

$$INC_{it} = \alpha + \beta \underbrace{EDU_{it}}_{\text{可观测随时间变化}} + \gamma \underbrace{GENDER_i}_{\text{可观测不随时间变化}} + \theta \underbrace{TALENT_i}_{\text{不可观测不随时间变化}} + \varphi \underbrace{LUCK_{it}}_{\text{不可观测随时间变化}}$$

- 我们把所有不可观测的因素都归于干扰项，因此上式对应的简单回归方程为

$$INC_{it} = \alpha + \beta EDU_{it} + \gamma GENDER_i + e_{it} \tag{1}$$

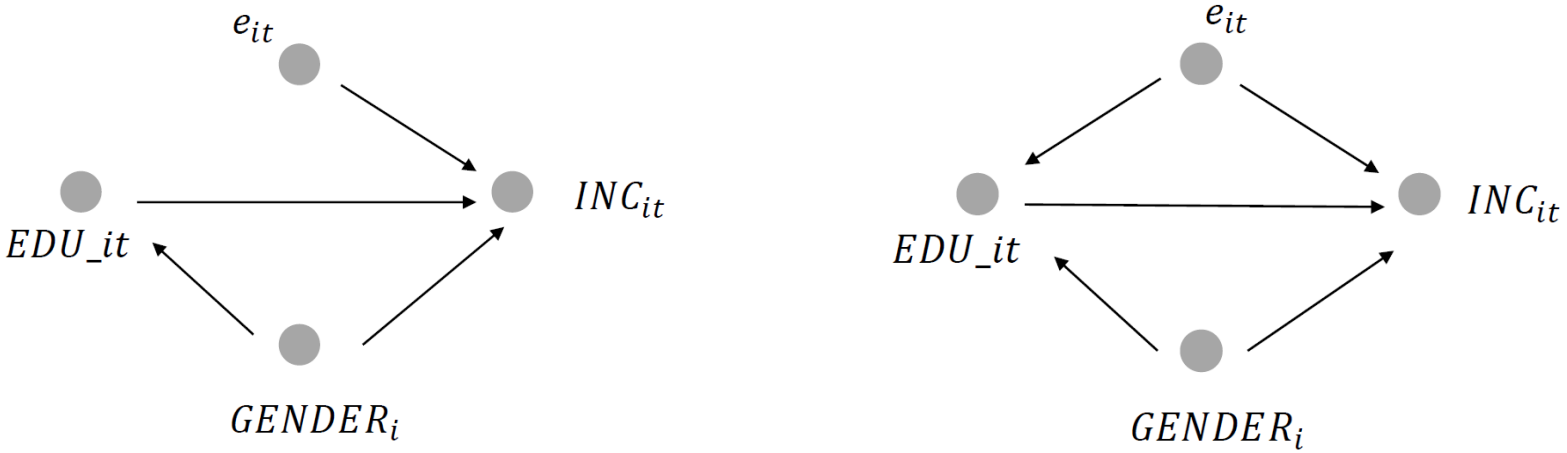
- 如果要得到 β 的正确估计，需要教育与干扰项不相关，即教育与个人先天能力以及运气都不相关。这显然是一个很强的假设条件。
- 面板数据允许我们将干扰项中不可观测且不随时间变化的因素从干扰项中分离出来。$e_{it} = \alpha_i + \mu_{it}$。其中，$\alpha_i$ 是个体不可观测且不随时间变化的因素为（个体效应），μ_{it} 是个体不可观测且随时间变化的因素。

面板数据模型因果关系分析的直观理解

- 那么，上式可以改写为：

$$INC_{it} = \alpha + \beta EDU_{it} + \gamma GENDER_i + \alpha_i + \mu_{it} \tag{2}$$

- 此时，要正确估计出 β，只需要满足教育和运气不相关即可。这个条件显然要更合理一些。
- 我们通过变量因果路径来进一步直观理解。



- 左图是我们希望的理想情况，即不存在混淆路径。右图是存在混淆变量的因果路径图。

面板数据模型因果关系分析的直观理解

- 那么，在使用包含个体效应模型（方程 2）的情况下，下图显示了其变量的路径图。

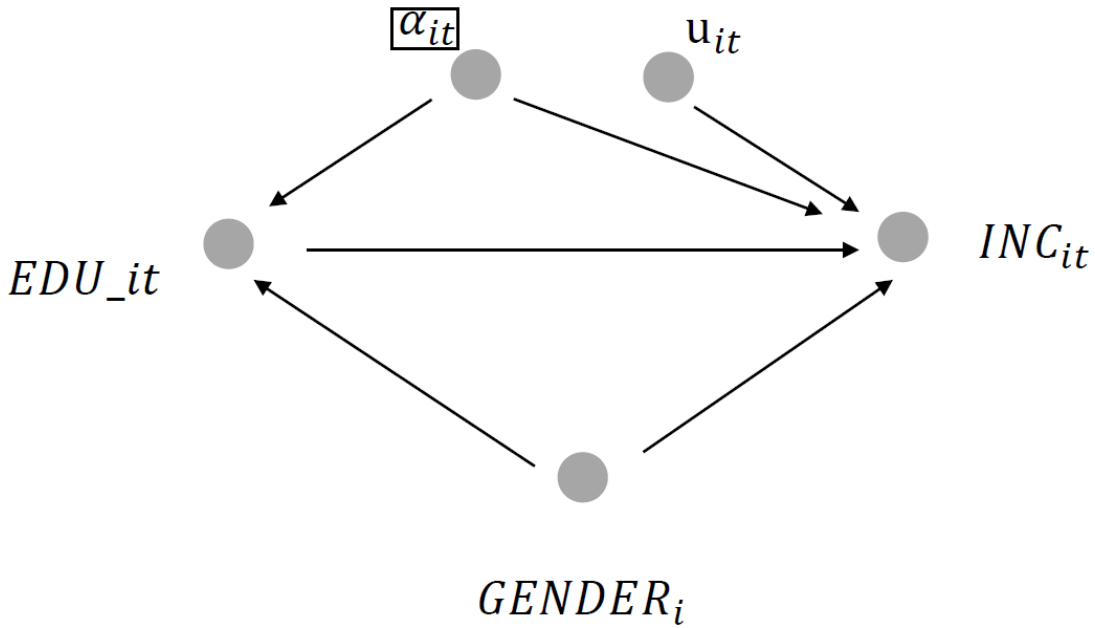


Figure: 面板数据的变量路径图

- 由上图可以看到，我们可以利用个体效应模型，将不可观测且不随时间变化的变量“控制住”。这在推断因果关系的有效性上前进了一步。

面板数据模型因果关系分析的直观理解

- **面板数据的估计策略**

$$INC_{it} = \alpha + \beta EDU_{it} + \gamma GENDER_i + \alpha_i + \mu_{it}$$

- 面板数据的三种模型都需要假设不可观测随时间变化的 μ_{it} 与可观测解释变量不相关。这三种模型的关键区别：对个体不可观测且不随时间变化的变量 α_i 的假设。它们分别是：混合截面数据模型；随机效应模型；固定效应模型。

1 混合截面数据模型：将面板数据看成是截面数据进行混合回归，要求样本中每个个体拥有完全相同的回归方程。即假设 α_i 不存在。但缺点是忽略了个体不可观测的异质性，而该异质性可能与解释变量相关，导致估计不一致。

2 随机效应模型：即假设 α_i 存在，但与可观测变量不相关。即个体效应是从某个分布中随机抽取，与其他解释变量无关。但是该假设在实际运用中通常是不成立的。（随机效应模型估计一般是采用 GLS 或 FGLS 方法来进行估计）

面板数据模型因果关系分析的直观理解

3 固定效应模型：假定个体的回归方程拥有相同的斜率，但可有不同的截距项，以捕捉异质性。即假设 α_i 存在，并与可观测变量相关。（见下图）

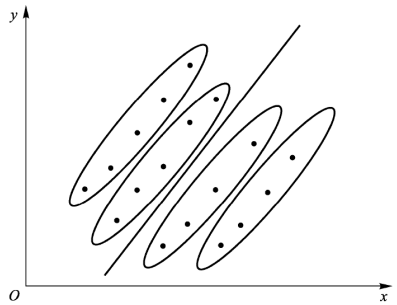


Figure: 面板数据中不同个体的截距项可以不同

- 对于固定效应模型，建议使用组内 R^2，对于随机效应模型，使用组内或组间 R^2 都可。
- 非平衡面板数据，固定效应模型和随机效应模型都适用。

固定效应模型估计方法

- **固定效应模型的一般形式**

$$Y_{it} = \underbrace{\boldsymbol{X'_{it}}}_{\text{可观测随时间变化}} \boldsymbol{\beta} + \underbrace{\boldsymbol{Z'_{i}}}_{\text{可观测不随时间变化}} \boldsymbol{\gamma} + \underbrace{\alpha_i}_{\text{不可观测不随时间变化}} + \underbrace{\mu_{it}}_{\text{不可观测随时间变化}} \tag{3}$$

$$E(\alpha_i|X_{it}, Z_i) \neq 0, E(\mu_{it}|X_{it}, Z_i, \alpha_i) = 0$$

- 这个模型里，我们不需要加入共同截距项，因为个体固定效应意味着每个个体都有一个单独的截距项。
- 固定效应模型由以下几种估计方法：个体内差分估计法；最小二乘虚拟变量估计法；一阶差分估计法。

固定效应模型估计方法

- **个体内差分估计法**

- 第一步，对每个个体的变量 $(Y_{it}, X_{it}, Z_i, \alpha_i, \mu_{it})$ 取个体平均值，

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it}, \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it}, Z_i = \frac{1}{T}\sum_{t=1}^{T} Z_i$$

$$\alpha_i = \frac{1}{T}\sum_{t=1}^{T} \alpha_i, \bar{\mu}_i = \frac{1}{T}\sum_{t=1}^{T} \mu_{it}$$

可得：

$$\bar{Y}_i = \bar{X}_i'\beta + Z'_i\gamma + \alpha_i + \bar{\mu}_i \tag{4}$$

- 第二步，（3）式减去（4）式：

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + (Z_i - Z_i)'\gamma + (\alpha_i - \alpha_i) + (\mu_{it} - \bar{\mu}_i)$$

$$Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)'\beta + (\mu_{it} - \bar{\mu}_i) \tag{5}$$

$$\tilde{Y}_{it} = \tilde{X}'_{it}\beta + \tilde{\mu}_{it}$$

固定效应模型估计方法

- 上式中不存在造成与解释变量相关的固定效应，$Cov(\tilde{X}'_{it}, \tilde{\mu}_{it}) = 0$，因此可以用 OLS 来对方程（5）估计：

$$\hat{\boldsymbol{\beta}}^{\boldsymbol{within}} = (\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}})^{-1}(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{Y}}) \tag{6}$$

  其中，$\hat{\boldsymbol{\beta}}^{\boldsymbol{within}}$ 也被称为个体内差分估计量、固定效应估计量。

- 这个方法依赖于 X 和 Y 在个体内变化的信息来估计它们的因果关系。经过个体内转换的数据不包含个体固定效应。

- 这个方法的代价是，通过个体内转换，不仅去除了不可观测的固定效应，同时也去除了可观测的不随时间变化的 Z。这也意味着可观测且不随时间变化的 Z 因为不存在个体内变化，其系数无法被估计。

固定效应模型估计方法

- **最小二乘虚拟变量估计法，**LSDV

- 用 N 个个体的虚拟变量来估计固定效应模型。其中，对于个体 i, 虚拟变量定义为，如果是个体 i, 那么 $D_i = 1$；如果是其他个体, 那么 $D_i = 0$.

$$Y_{it} = X'_{it}\beta + Z'_{i}\gamma + \alpha_1 D_1 + \alpha_2 D_2 + \cdots + \alpha_N D_N + \mu_{it} \qquad (7)$$

Table: 包含个体虚拟变量的面板数据

ID	YEAR	INC	EDU	GENDER	D1	D2
1	2017	800	3	1	1	0
1	2018	1000	4	1	1	0
1	2019	1200	5	1	1	0
2	2017	1200	5	0	0	1
2	2018	1250	6	0	0	1
2	2019	1300	7	0	0	1

固定效应模型估计方法

- 由上表可以看出，可观测且不随时间变化的 GENDER 与个体虚拟变量存在完全共线性问题。因此，上式无法估计出 GENDER 的系数。所以，通常固定效应模型不包含可观测且不随时间变化的变量。

$$Y_{it} = X'_{it}\beta + \sum_{i=1}^{N} d_i D_i + \mu_{it} \tag{8}$$

这里，$d_i = Z'_i\gamma + \alpha_i$。

- 通常，变量 Z 的作用被个体固定效应“吸收”。
- LSDV 法和个体差分估计法得到的结果是完全一样的。回归分解法告诉我们，如果要估计 X 的系数，有两种等价的方法。一种方法是将 Y 对 X 和 Z 一起回归，直接得到估计系数 $\hat{\beta}_1$。另一种方法是使用两步法。第一步先将 Y 和 X 分别对 Z 进行回归，得到 Y 和 X 的残差值；第二步将 Y 的残差值和 X 的残差值进行回归，得到 X 的残差值的回归系数 $\hat{\beta}_1$。
- LSDV 方法的缺点是，如果 N 很大，则方程中引入很多虚拟变量，这会使得估计的计算量变得很大。

固定效应模型估计方法

- **一阶差分估计法**

- 一阶差分法和个体内差分估计法的原理类似。一阶差分估计法通过对每个个体前后两期做差分，达到去掉个体固定效应目的。

$$Y_{it} = X'_{it}\beta + d_i + \mu_{it} \tag{9}$$

$$Y_{it-1} = X'_{it-1}\beta + d_i + \mu_{it-1} \tag{10}$$

两式相减，得到：

$$\Delta Y_{it} = \Delta X'_{it}\beta + \Delta\mu_{it} \tag{11}$$

- 由于 $\Delta X'_{it}$ 和 $\Delta\mu_{it}$ 不相关，故使用 OLS 方法估计得到的 $\hat{\beta}$ 是一致估计量。

- 当 T=2，一阶差分估计法与个体内差分估计法的结果是一样的。当 T>2，一阶茶法估计法与个体内差分估计法的结果是不一样的，二者会因为采用误差而有所差异。

固定效应模型估计方法

- **时间固定效应的引入**

- 个体固定效应解决了不随时间而变但随个体而异的遗漏变量问题。通过引入时间固定效应，可以控制在同一时间不随个体变化的变量（比如每年的宏观经济因素）。

$$Y_{it} = X'_{it}\beta + d_i + T_t + \mu_{it}$$

- 其中，T_t 是时间固定效应。由于上述方程既考虑个体固定效应，又考虑时间固定效应，故被称为双向固定效应模型。

- 由于时间点通常不多，可以引入 T-1 个时间虚拟变量在回归方程中，避免共线性。

$$Y_{it} = X'_{it}\beta + d_i + \sum_{t=2}^{T} \gamma T_t + \mu_{it} \tag{12}$$

其中，如果观测点是在时间 t, 那么 $T_t = 1$；否则 $T_t = 0$。

固定效应模型估计方法

- **时间固定效应的引入**

- 有时为了节省参数（比如，时间维度 T 较大），可引入时间趋势项，以替代上式（T-1）个时间虚拟变量。

$$Y_{it} = X'_{it}\beta + d_i + \gamma t + \mu_{it}$$

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

- 在使用家庭（企业）层面，省（市、县）、行业层面的数据时，还需要同时控制家庭固定效应、省份固定效应和行业固定效应等。

固定效应模型估计方法

- **个体效应估计残差**

- 个体效应 d_i 的估计量为，

$$\hat{d}_i = \bar{Y}_i - \bar{X}_i'\beta \tag{13}$$

- $\hat{d}_i$ 是 $d_i = Z'_i\gamma + \alpha_i$ 的一致估计量。只有当不存在 Z 时，$\hat{d}_i$ 才是 α_i 的一致估计量。

- 如果想估计 Z 的回归系数。当 Z_i 和 α_i 不相关（这是前提），将 $\hat{d}_i$ 对 Z_i 进行回归，可以得到 γ 一致估计量。

- 估计的固定效应 $\hat{d}_i$ 只有在时间维度 T 趋向无穷时，才是一致估计量。这是因为当个体维度 N 增加时，我们有很多的 d_i 需要估计。只有当 T 增加时，我们才有更多的信息去估计每个固定效应。但是，通常面板数据的 T 都不可能足够大，因此通常 $\hat{d}_i$ 是不一致的。这就是所谓的“偶发参数问题”。所以，我们一般并不计算和使用个体效应估计量 $\hat{d}_i$。

动态面板的 GMM 估计方法

- 动态面板数据模型：在面板数据中考虑被解释变量的动态特征。
- 由于被解释变量的滞后项也进入回归方程，个体固定效应会导致 OLS 回归产生偏误和不一致性。
- 为了克服 OLS 估计的问题，需要引入工具变量：在动态面板数据模型中，最常用的工具变量是被解释变量和解释变量的滞后及差分滞后项。引入这类工具变量后，可利用 GMM 的一般框架来进行估计。

动态面板的 GMM 估计方法

- **动态面板数据模型的基本形式**

$$Y_{it} = \phi Y_{it-1} + X'_{it}\beta + \mu_i + \varepsilon_{it} \tag{14}$$

其中，β、X 是向量，ε_{it} 表示残差项且与所有解释变量无当期相关性。

- 下面会进一步放松 X_{it} 与 ε_{it} 的相关性假设。
- 个体固定效应会导致动态面板回归产生内生性问题：回归变量与残差项 $\mu_i + \varepsilon_{it}$ 相关性不为 0。
- 当面板数据模型具有动态特征时，静态面板的简单变换方法无法消除固定效应的影响，因而 OLS 估计不可行。

动态面板的 GMM 估计方法

- 以差分方法为例：对回归方程两边取差分后得：

$$\Delta Y_{it} = \phi \Delta Y_{it-1} + \Delta X'_{it} \beta + \Delta \varepsilon_{it}$$

  但 $Cov(\Delta Y_{it}, \Delta \varepsilon_{it}) \neq 0$，$\varepsilon_{it-1}$ 对 Y_{it-1} 有影响。

- 组内平均的变换带来的问题更大。

动态面板的 GMM 估计方法

- **基本解决思路：滞后项作为工具变量**

- 差分法在动态面板下无法克服固定效应带来的问题：根源在于被解释变量的动态性 $\Rightarrow$ t 期回归变量 Y_{it-1} 与残差差分项中 ε_{it-1} 的相关性。
- 如果能找到一个变量，与差分方程中的回归变量相关，但与残差无关，则可用作工具变量。
- 最简单的工具变量：Y_{it-2}，与 ΔY_{it-1} 相关，但与 ε_{it} 无关！（Anderson and Hsiao,1981）

动态面板的 GMM 估计方法

- **基本假设 1：** ε_{it} 没有序列相关性

- Arrelano and Bond(1991,RES) 提出把所有的滞后项全部引入为工具变量，使用 GMM 方法进行估计。由于这一方法的基础是对差分方程进行估计，故称为差分 GMM 模型。

- 这样 $\{Y_{it-3}, Y_{it-4}, \cdots\}$ 都是有效的工具变量，可以大幅度提高估计效应。

- 差分 GMM 的问题

1 因为取了差分，所有纯粹截面作用无法估计。

2 差分可能会降低信噪比 $\Rightarrow$ 水平值时间、截面差异的信息都被弱化。

3 如果被解释变量和解释变量的一阶自相关很接近 1，则可能存在弱工具变量问题。如 $\phi \approx 1$, 那么 $\Delta Y_{it} = Y_{it} - Y_{it-1}$J 几乎完全由回归方程中的余项决定，与 Y_{it-1} 的相关系数会很小。

动态面板的 GMM 估计方法

- 针对上述问题，特别是高自相关性带来的弱工具变量问题，Arrelano and Bover(1995) 提出了另外一个选择工具变量的方法：水平回归与差分滞后。
- 回到动态面板水平回归方程：

$$Y_{it} = \phi Y_{it-1} + X'_{it}\beta + \mu_i + \varepsilon_{it} \tag{15}$$

- 如果被解释变量差分滞后项 ΔY_{it-1} 与固定效应没有关系，那么 ΔY_{it-1} 也可以作为工具变量对水平方程进行 GMM 估计！

- **基本假设** 2:$Cov(\Delta Y_{it-1}, \mu_i) = 0$

- 在此假设下，结合差分 GMM 和水平 GMM，将差分方程和水平方程作为一个方程系统，进行 GMM 估计。这一方法被 Blundell and Bond(1998) 称为系统 GMM。

动态面板的 GMM 估计方法

- Blundell and Bond 指出了基本假设 2 的实质：Y_{it} 关于其长期均值的暂时偏离与 μ_i 无关。换言之，个体的初始偏离水平与长期趋势无关。（Blundell and Bond，1998；Roodman,2009）
- 优点：提高估计效率；可以估计不随时间变化的变量的系数；当样本自相关性比较高时，系统 GMM 的有限样本偏差要比差分 GMM 好。
- 缺点：当 $\{\Delta Y_{it-1}, \Delta Y_{it-2}, \cdots\}$ 与个体效应相关时，无法使用这一模型。

动态面板的 GMM 估计方法

- **模型设定检验**

1 检验残差项 ε_{it} 的自相关性

- 基本假设 1 使得 $\Delta\varepsilon_{it}$ 不会具有 2 阶及以上自相关。

2 过度识别检验（Sargan 检验）

- 原假设为不存在过度识别
- 当拒绝原假设时（p 值较小），需要考虑缩减工具变量的数量或者是增多被解释变量滞后期的设定。

动态面板的 GMM 估计方法

- 前面讨论中我们假定解释变量完全外生。
- 上述假设可以放松：允许存在前定前面和内生变量。
- 前定变量：ω_{it} 与 ε_{it} 不相关，但是与 ε_{it-1} 及更高阶滞后相关。
- 内生变量：ω_{it} 与 ε_{it} 相关及更高阶滞后相关。

动态面板的 GMM 估计方法

- **如何判断一个经济模型中的前定变量，内生变量和外生变量？**
- 数据：595 名美国工人 1976-1982 年有关工资（短面板）
- 被解释变量：lwage(工资对数)
- 解释变量：
- 外生变量:occ(是否是蓝领工人),south（是否在南方）,smsa（是否在大城市）,ind（是否在制造业）
- 前定变量：wks(已工作周数)
- 内生变量：ms（婚否）,union（是否由工会合同确定工资）
- 不随时间变化的变量：ed(教育年限),fem（是否女性）,blk（是否黑人）。由于对原模型进行了差分处理，这些变量将无法被估计。

动态面板的 GMM 估计方法

- 动态面板数据模型

$$
\begin{aligned}
lwage_{it} = \alpha + & \underbrace{\rho_1 lwage_{it-1} + \rho_2 lwage_{it-2}}_{\text{被解释变量滞后项}} \\
& + \underbrace{\beta_1 occ_{it} + \beta_2 south_{it} + \beta_3 smsa_{it} + \beta_4 ind_{it}}_{\text{外生变量}} \\
& + \underbrace{\beta_5 wks_{it} + \beta_6 wks_{it-1}}_{\text{前定变量}} \\
& + \underbrace{\beta_7 ms_{it} + \beta_1 union_{it}}_{\text{内生变量}} + \mu_i + \varepsilon_{it}
\end{aligned}
$$

-

$$
\hat{d}_i = \bar{Y}_i - \bar{X}_i' \beta \tag{16}
$$

动态面板的 GMM 估计方法

- 矩估计：利用样本矩来估计总体相应的参数。
- 常见的有一阶矩（期望），二阶矩（方差）。根据大数定理，样本矩去取代总体矩。
- 广义矩估计就是来解决过度识别问题。OLS、2SLS 都是特殊的广义矩估计。
- GMM 的优点：把复杂的统计过程抽象化成为一个 (看似) 简单的过程：找矩条件。只要你能找到矩条件，你就能估计。
- GMM 认为不能使所有方程成立（过度识别），但是可以让方程尽可能的接近于 0。

动态面板的 GMM 估计方法

- **球形扰动项假设下，2SLS 是最有效率的。扰动项存在异方差或自相关，广义矩估计（Generalized Method of Momens,GMM）更有效**
  - 总体矩条件

$$E(g_i) = E(z_i, \varepsilon_i) = 0$$

  - 样本矩条件

$$g_n(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x'_i\hat{\beta}) = 0$$

未知参数 $\hat{\beta}$ 有 K 个，方程个数有 L 个（z_i 的维度）

$L < K$，不可识别；$L = K$, 恰好识别，有唯一解；$L > K$, 过度识别，无解，传统的矩估计方法求出。

动态面板的 GMM 估计方法

- 利用“权重矩阵”W 来构造二次型
- $\hat{W}$ 是 $L \times L$ 维对称正定矩阵，$p\lim \hat{W} = W$,W 为非随机的对称正定矩阵。
- 最小化目标函数

$$\min_{\hat{\beta}} J(\hat{\beta}, \hat{W}) = n[g_n(\hat{\beta})]' \hat{W} [g_n(\hat{\beta})]$$
$$\hat{\beta}_{GMM}(\hat{W}) = \arg\min_{\hat{\beta}} J(\hat{\beta}, \hat{W})$$

- 最小化问题的解（过程略）

$$\hat{\beta}_{GMM}(\hat{W}) = ({Q'}_{zx} \hat{W} Q_{zx})^{-1} {Q'}_{zx} \hat{W} Q_{zy}$$

其中，

$$Q_{zx} = \frac{1}{n} \sum_i Z_i {X'}_i, Q_{zy} = \frac{1}{n} \sum_i {Z'}_i y_i$$

动态面板的 GMM 估计方法

- **恰好识别下，Q_{zx} 为方阵，GMM 还原为工具变量法**
- 最优 GMM:$\hat{W} = \hat{Q}^{-1}$ 是使 $Avar(\hat{\beta}_{GMM})$ 最小化的最优权重矩阵，使用 $\hat{Q}^{-1}$ 为权重矩阵的 GMM 估计量为“最优 GMM”。
- 估计步骤
- 使用 2SLS，得到残差，计算 $\hat{Q} = \frac{1}{n}\sum_{i=1}^{n}(e_i)^2 z_i z'_i$
- 最小化 $J(\hat{\beta}, \hat{Q}^{-1})$, 得到 $\hat{\beta}_{GMM}(\hat{Q}^{-1})$
- 迭代法（实际操作中常用）
- 用第二步所得到的的残差再来计算 $\hat{Q}$, 然后再来计算 $\hat{\beta}_{GMM}(\hat{Q}^{-1})$
- 迭代法（实际操作中常用），直到估计值收敛。

面板数据模型常见问题

- **固定效应模型 VS 随机效应模型**

- 判断是选择 FE 模型还是 RE 模型，取决于个体效应 α_i 是否与解释变量相关。

	$cov(X_{it}, \alpha_i) = 0$	$cov(X_{it}, \alpha_i) \neq 0$
随机效应模型(RE)	$\hat{\beta}^{RE}$一致，有效	$\hat{\beta}^{RE}$不一致
固定效应模型(FE)	$\hat{\beta}^{FE}$一致，不有效	$\hat{\beta}^{FE}$一致，可能有效

- 一般，我们采用 Hausman 检验来判断使用哪一个模型。Hausman 检验的原假设是：α_i 与 X_{it} 不相关。

$$H = (\hat{\beta}^{FE} - \hat{\beta}^{RE})'[Var(\hat{\beta}^{FE}) - Var(\hat{\beta}^{RE})]^{-1} \times (\hat{\beta}^{FE} - \hat{\beta}^{RE}) \rightarrow \chi^2(K) \tag{17}$$

K 是自由度，为模型中自变量的取值个数。

面板数据模型常见问题

- 如果原假设成立，$\hat{\beta}^{FE}$ 和 $\hat{\beta}^{RE}$ 都是一致的，$(\hat{\beta}^{FE} - \hat{\beta}^{RE})$ 接近于 0，因此 H 值应该接近于 0，则说明 $(\hat{\beta}^{FE} - \hat{\beta}^{RE})$ 差异较大，也就是说 RE 模型是不一致的，应该选用 FE 模型。
- 传统的 Hausman 检验不适用于异方差的情形，须使用异方差稳健的 Hausman 检验。
- 在实际应用中，RE 模型要求的条件很难满足。因此多数情况下，我们并不需要 Hausman 检验，而常常直接使用 FE 模型。

面板数据模型常见问题

- 有些变量在使用固定效应模型后，系数大小和显著性发生很大变化关键在于理解这些变量的信息来源。
- 使用固定效应模型后，通常估计系数的方差会变大。这是因为，为避免不可观测的个体效应造成的遗漏变量误差，固定效应模型估计依赖于个体内信息来估计回归系数。这造成个体间平均差异的信息没有被用来估计 X 和 Y 的关系，由于使用的信息量变少了，就造成估计系数方差变大。

面板数据模型常见问题

- **使用固定效应模型后，有些变量系数变为不显著，是否意味着该解释变量和被解释变量没有因果关系？**

- 一个原因是，变量 X 对 Y 确实没有因果影响（因果关系系数 =0）。在没有使用固定效应模型，存在遗漏变量偏误，造成 Y 对 X 的回归系数显著不等于 0。在使用固定效应模型控制了遗漏变量的误差后，Y 对 X 的回归系数和 0 不显著差异。
- 另一个原因是，变量 X 对 Y 有因果影响。但变量 X 的个体内变化太小，控制了固定效应后，由于信息不够，造成估计系数方差太大而导致 X 的系数不显著，但这并不代表 X 对 Y 没有因果影响。

面板数据模型常见问题

- 例如，一个 10 年面板数据，大学生与高中生相比，在 10 年中一直保持较高的收入，但每个人的学历在 10 年里很少发生变化。使用固定效应模型，由于很少人的学历有个体内变化，因此用个体内学历变化去估计其对收入的影响很难得到显著的结果。但实际情况是，学历和收入是有因果关系的，只是它们的关系并不能由个体内变化反映出来，而是体现在个体间差异上，
- 因此，固定效应模型是有局限性的。在使用面板数据前，对每个变量，尤其是所关注的变量的信息来源要有充分的了解。

参考文献：

陈强. 计量经济学及 Stata 应用 [M]. 高等教育出版社，2015