面板回归基础

报告人：刘岩

武汉大学经管学院金融系

2017 年 9 月

本讲内容

面板回归的用处

- 与普通截面、时间序列回归相比，面板回归利用面板数据额外的维度与增加的样本量，提高估计的效率（efficiency ）与推断的功效（power）。
- 此外，面板回归还提供了简单的方法，处理变量不可观测性带来的遗漏变量（omitted variable）问题，从而保证估计的一致性（consistency）。

本节内容

面板数据

- 面板数据的每个样本观测值（sample observation）至少有截面 i 和时间 t 两个维度：如地区-年份，企业-年份。
  - 允许更高的维度：(i, j, t)，如地区-银行-年份。
- 截面单位个数：$i = 1, \ldots, N$；时间单位个数：$t = 1, \ldots, T$。
- 面板数据分类：大 N 小 T，小 N 大 T，大 N 大 T。
  - 大部分研究中碰到的情况都是大 N 小 T：中国地市面板 $N \approx 300, T \approx 15$；中国银行面板 $N \approx 150, T \approx 10$；跨国面板 $N \approx 180, T \approx 30$。

面板数据分类

- 平衡面板（balanced panel）
  每个截面单位的样本时间长度均为 T，样本总数为 $N \times T$。
- 非平衡面板（unbalanced panel）
  截面 i 的样本时间长度为 $T_i \leq T$，样本总数为 $\sum_{i=1}^{N} T_i$。
- 大部分面板回归的方法对平衡面板和非平衡面板均适用。

面板回归模型

- 给定一组面板数据 $\{y_{it}, \boldsymbol{x}_{it}\}$，$y_{it}$ 是被解释变量（dependent variable），$\boldsymbol{x}_{it}$ 是解释变/向量（independent/explanatory variable, regressor）。
- 研究中关心的是 y_{it} 与 $\boldsymbol{x}_{it}$ 间的关系，或者说，$\boldsymbol{x}_{it}$ 如何影响 y_{it}；一般而言，这一关系可以表示为 $y_{it} = f(\boldsymbol{x}_{it})$。
- 研究中最常用的面板模型：线性面板回归模型

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + v_{it},$$

其中 $\boldsymbol{\beta}$ 是回归系数（向量），v_{it} 表示均值为 0 的残差项。

- 识别（估计）$\boldsymbol{\beta}$：同时利用 y_{it} 和 $\boldsymbol{x}_{it}$ 在截面和时间两个维度上的变化。

个体效应与时间效应

- 残差项 v_{it} 一般可以分解为三部分：

$$v_{it} = u_i + \delta_t + \varepsilon_{it}.$$

- 上式中，u_i 表示个体效应（individual effect），δ_t 表示时间效应（time effect），ε_{it} 通常假设为在 i 和 t 两个维度均相互独立且与其他变量 $(\boldsymbol{x}_{it}, u_i, \delta_t)$ 相独立。
- 引入个体效应和时间效应，可以非常方便的控制无法观测到的个体特征和时间变化对 y_{it} 的影响。

面板回归模型与 OLS 估计

- 通常情况下，OLS（ordinary least square）方法是估计回归系数 $\boldsymbol{\beta}$ 的最简单方法。
- OLS 估计值 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 具有一致性的前提假设：$\boldsymbol{x}_{it}$ 与 v_{it} 不具有相关性。
  - 回归模型脱胎于经济理论，因此 $\boldsymbol{\beta}$ 的值本质上由理论确定。
  - 由于样本是随机的，$\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 也是一个随机变量。
  - $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 具有一致性：当样本足够大时，其概率极限逼近理论值 $\boldsymbol{\beta}$，
$$\operatorname*{plim}_{NT\to\infty} \hat{\boldsymbol{\beta}}_{\text{OLS}} = \boldsymbol{\beta}.$$
  - 一致性是参数估计的最基本要求：理论 $\to$ 数据 $\to$ 理论。

单一解释变量的例子

考虑单一解释变量 x_{it} 的情形，此时有

$$\hat{\beta}_{\text{OLS}} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}y_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2} = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(x_{it}\beta + v_{it})}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2}$$
$$= \beta + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}v_{it}}{\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2} = \beta + \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}v_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2}.$$

通常假定 $\text{plim}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2$ 是一个常数，则 $\hat{\beta}_{\text{OLS}}$ 的一致性取决于 $\text{plim}\ \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}v_{it} = \text{cov}(x_{it}, v_{it}) = 0$ 是否成立：

关键在于 $\text{cov}(x_{it}, u_i) = 0$ 及 $\text{cov}(x_{it}, \delta_t) = 0$。

此外，个体/时间效应还可影响 $\hat{\beta}_{\text{OLS}}$ 的方差（标准误）。

OLS 估计下引入虚拟变量

- 绝大多数面板数据都是大 N 小 T 型，在此情况下，时间效应不会带来任何问题。
  - 由于此情形下一般假设 T 固定，$N \to \infty$，因此可以设置固定数量的时间虚拟变量 D_t，直接估计出 δ_t，避免其与 x_{it} 之间的相关性干扰 OLS 估计。
- 对于个体效应，无法使用虚拟变量来解决问题：D_i 的数目随 N 的增大而增大。

个体效应的问题

- 通常情况下无法保证 $\mathrm{cov}(x_{it}, u_i) = 0$，因此 OLS 估计很可能出现不一致。
- 此外，即便有 $\mathrm{cov}(x_{it}, u_i) = 0$，$\hat{\beta}_{\mathrm{OLS}}$ 的标准误（standard error）——即 $\hat{\beta}_{\mathrm{OLS}}$ 作为随机变量的标准差——也需要考虑到个体效应 u_i 的影响。

本节内容

随机效应的界定

- 下面只考虑个体效应，则回归模型残差项为 $v_{it} = u_i + \varepsilon_{it}$。
- 若 $\text{cov}(\boldsymbol{x}_{it}, u_i) = 0$，则 u_i 称为**随机效应**。
- 在此情况下，面板回归模型的 OLS 估计 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 具有一致性。
- 问题是：如何对 $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ 进行统计推断？
  - 此种情形下最好的参考文献：Petersen 2009 RFS "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches."

随机效应模型 OLS 估计的参数推断

继续考虑单一解释变量的例子，此时 $\hat{\beta}_{\text{OLS}}$ 的渐进方差（asymptotic variance）可以写作

$$
\begin{aligned}
&\text{AV}[\hat{\beta}_{\text{OLS}} - \beta] \\
= &\operatorname*{plim}_{N\to\infty} \frac{1}{N^2} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} x_{it} v_{it} \right)^2 \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^2 \right)^{-2} \\
= &\operatorname*{plim}_{N\to\infty} \frac{1}{N^2} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} x_{it}^2 v_{it}^2 + 2 \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} x_{it} x_{is} v_{it} v_{is} \right) \\
&\times \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} x_{it}^2 \right)^{-2}.
\end{aligned}
$$

残差 v_{it} 自然具有序列相关性：$\text{cov}(v_{it}, v_{is}) \neq 0$。

前述标准误的说明

- 此处没有任何附加的同方差（homoskedasticity）假设；允许异方差（heteroskedasticity），$\mathbb{E}[v_{it}] \neq \mathbb{E}[v_{jt}]$。
  - 允许异方差的标准误估计称为异方差稳健标准误，或 White 稳健标准误。
- 这里的确用到的假设：v_{it}，实质是 u_i，在截面上相互独立。
  - 上页第二个等号用到这个假设。

聚类标准误

- 这一类标准误的估计，又称为聚类标准误（clustered standard error），或截面个体层面聚类标准误（s.e. by cluster at individual level）。
  - 也可以不在截面个体层面上进行聚类，比如地区-银行-年份面板，可以在地区-银行层面聚类（相当于上面的截面个体具体），也可以在地区层面聚类：此时允许同一地区不同银行残差项相关。
- 聚类标准误具有很好的稳健性；用此标准误进行的统计推断最为准确——这类标准误又称为面板稳健标准误（panel robust s.e.）
- **关键**：聚类标准误可以捕捉残差序列相关性 $\mathrm{cov}(v_{it}, v_{is})$ 对 $\hat{\beta}_{\mathrm{OLS}}$ 标准误的影响。

STATA 的处理

- STATA 的标准处理：如无特别说明，则在同方差假设下使用 GLS 估计。
  - 同方差：$\mathbb{E}[u_i^2] = \sigma_u^2, \mathbb{E}[\varepsilon_{it}^2] = \sigma_\varepsilon^2$；此时 GLS 估计值等价于对数据进行变换后的 OLS 估计：

$$y_{it} - \lambda \bar{y}_i = (\boldsymbol{x}_{it} - \lambda \bar{\boldsymbol{x}}_i)\boldsymbol{\beta} + (1-\lambda)u_i + \varepsilon_{it} - \lambda \bar{\varepsilon}_i,$$

  其中 $\lambda = 1 - \sigma_\varepsilon / \sqrt{T\sigma_u^2 + \sigma_\varepsilon}$.

- 可以在标准误估计中选择 robust 选项；`vce(robust)`；该选项等价于 `vce(cluster)`。

本节内容

固定效应的处理

- 当 $\operatorname{cov}(\boldsymbol{x}_{it}, u_i) \neq 0$ 时，称为个体固定效应。
- 此时对原始面板回归的 OLS 估计会有不一致性：以单一解释变量为例，

$$\hat{\beta}_{\mathrm{OLS}} = \beta + \frac{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}v_{it}}{\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}^2} \neq \beta,$$

原因在于

$$\begin{aligned} &\operatorname{plim} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}v_{it} \\ = &\operatorname{plim} \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}(u_i + \varepsilon_{it}) \neq 0. \end{aligned}$$

个体固定效应模型的 OLS 估计方法

- 此种情况下，为避免 u_i 引起的 OLS 估计偏误问题，可以通过简单的差分方法去除固定效应：

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}' \boldsymbol{\beta} + \Delta \varepsilon_{it},$$

如此可以使用 OLS 进行估计。

- 也可以使用组内均值（within group）方法去除固定效应：令 $\bar{y}_i = \frac{1}{T}\sum_t y_{it}$，$\bar{\boldsymbol{x}}_i = \frac{1}{T}\sum_t \boldsymbol{x}_{it}$，$\bar{\varepsilon}_i = \frac{1}{T}\sum_t \varepsilon_{it}$，则有

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

亦可使用 OLS 估计。

固定效应模型的推断和问题

- 固定效应模型不论选择差分估计还是组内估计，OLS 估计值 $\hat{\beta}_{\text{OLS}}$ 都受到残差序列自相关性的影响，因此都可以使用聚类标准误。
  - 差分估计里，残差序列为 $\varepsilon_{it} - \varepsilon_{it-1}$，存在一阶自相关性。
  - 组内估计里，残差序列为 $\varepsilon_{it} - \bar{\varepsilon}_i$，若 T 较小，那么共同因子 $\bar{\varepsilon}_i$ 的影响较大。
- 除此之外，还可以考虑异方差问题。
- 个体固定效应模型的 OLS 估计，最后主要利用的是个体观测值在时间上的变化，即组内变动（within variation），而对观测值的组间变动（between group variation）利用有限。

STATA 的处理

- STATA 默认使用组内估计方法对固定效应模型进行估计。
- 系数估计值的标准误计算：vce(robust) 和 vce(cluster) 等价，均给出面板稳健标准误——同时考虑了残差的异方差和序列相关性问题。
  - vce(cluster groupvar) 还可以指定按组类计算聚类标准误：如地区-银行-年度面板中，默认聚类是在地区-银行层面，如果选择按地区分组，那么标准误的计算可以考虑到同一地区内不同银行、不同年份间残差的相关性。

固定效应与随机效应模型对比

- 固定效应要求的假设要弱于随机效应：一般情况下很难有足够证据说明个体效应 u_i 和个体回归变量 $\boldsymbol{x}_{it}$ 不相关。
- 固定效应估计的问题：只能估计时变回归变量的系数；若 $x_{i1} = x_{iT}$，那么其对应系数估计不出来。此外，固定效应模型更偏回归变量时间变动的信息，对截面变动信息利用较少。
- 随机效应模型不存在上述困难。

Hausman 检验

- 文献中习惯使用 Hausman 检验来判断个体效应的类型：随机 vs. 固定。
- 该检验的原假设如下

$$H_0 : \mathrm{cov}(u_i, x_{it}) = 0.$$

若不拒绝原假设，选择随机效应；若否，选择固定效应。

- 在同方差假设下，随机效应估计 $\hat{\beta}_{\mathrm{RE}}$ 在原假设 H_0 下是一致有效估计，而固定效应估计 $\hat{\beta}_{\mathrm{FE}}$ 也是一致估计，因此大样本下 $\hat{\beta}_{\mathrm{RE}} \approx \hat{\beta}_{\mathrm{FE}}$。此时可构造 Hausman 检验统计量

$$(\hat{\beta}_{\mathrm{FE}} - \hat{\beta}_{\mathrm{RE}})' \mathrm{var}((\hat{\beta}_{\mathrm{FE}} - \hat{\beta}_{\mathrm{RE}}))^{-1} (\hat{\beta}_{\mathrm{FE}} - \hat{\beta}_{\mathrm{RE}}) \to \chi^2(k),$$

k 为 $\boldsymbol{x}_{it}$ 中随时间而变的回归变量个数。

Hausman 检验的问题

- Hausman 检验的第一个问题：需要同方差假设。如果模型本身是异方差，估计时也使用了稳健标准误，那么上述 Hausman 检验无效。——此时需要使用辅助回归方法，构造过度识别检验；STATA 命令为 `xtoverid`。详见陈强 (2014, pp. 269–270)。
- 此外，Hausman 检验是一个模型设定检验：不单纯是检验随机效应-固定效应，而是整个模型。回归变量设置不当等问题也会引起 Hausman 拒绝或接受原假设。因此，Hausman 检验只是个体效应类型选取的参考。