计量经济学

混合横截面和面板数据方法

张晨峰

2016年5月24日

华东理工大学商学院

6. 独立混合横截面和面板数据方法

主要内容

- 独立混合横截面
- 面板数据方法

6.1 独立混合横截面

独立混合横截面

- 它是在不同时点从一个大总体里进行随机抽样的结果。
- 它们都是由独立抽取的观测所构成。
- 它与一个随机样本的差异在于，在不同时点上对总体进行抽样很可能导致观测点不是同分布的。

面板数据

- 要收集面板数据，要在不同时间跟踪相同的个人、家庭、企业或城市或别的什么单位。
- 就面板数据的计量经济分析而言，我们不能假定，不同时点的观测是独立分布的。

独立混合横截面

- 使用独立混合横截面的一个理由是要加大样本容量。
- 但仅当因变量和某些自变量保持着不随时间而变量的关系时，混合才是有用的。
- 为了反映总体在不同时期会有不同的分布，可以允许截距在不同时期有不同的值。

6.1 独立混合横截面

TABLE 13.1 Determinants of Women's Fertility

Dependent Variable: *kids*

Independent Variables	Coefficients	Standard Errors
educ	−.128	.018
age	.532	.138
age^2	−.0058	.0016
black	1.076	.174
east	.217	.133
northcen	.363	.121
west	.198	.167
farm	−.053	.147
othrural	−.163	.175
town	.084	.124
smcity	.212	.160
y74	.268	.173
y76	−.097	.179
y78	−.069	.182
y80	−.071	.183
y82	−.522	.172
y84	−.545	.175
constant	−7.742	3.052
$n = 1{,}129$ $R^2 = .1295$ $\bar{R}^2 = .1162$		

6.1 独立混合横截面

A log(*wage*) equation (where *wage* is hourly wage) pooled across the years 1978 (the base year) and 1985 is

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot female + u, \quad [13.1]$$

Now, we use the data in CPS78_85.RAW to estimate the equation:

$$\begin{aligned} \log(wage) = {} & .459 + .118\, y85 + .0747\, educ + .0185\, y85 \cdot educ \\ & (.093)\ \ (.124)\ \ \ \ (.0067)\ \ \ \ (.0094) \\ & + .0296\, exper - .00040\, exper^2 + .202\, union \\ & \ \ (.0036)\ \ \ \ (.00008)\ \ \ \ (.030) \\ & - .317\, female + .085\, y85 \cdot female \\ & \ \ (.037)\ \ \ \ (.051) \\ n = {} & 1{,}084,\ R^2 = .426,\ \bar{R}^2 = .422. \end{aligned} \quad [13.2]$$

利用混合横截面做政策分析

当某些外生事件（常常是政府的政策改变）改变了个人、家庭、企业或城市运行的环境时，便产生了自然实验。一个自然实验总有一个不受政策变化影响的对照(control)组和一个被认为受政策变化影响的处理(treatment)组。为了控制好对照组和处理组之间的系统差异，我们需要两个年份的数据，一个在政策改变以前，另一个在政策改变之后。

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + \mu$$

其中，$\hat{\delta}_1$是倍差估计量

$$\hat{\delta}_1 = (\bar{y}_{2,T} - \bar{y}_{2,C}) - (\bar{y}_{1,T} - \bar{y}_{1,C})$$

6.1 独立混合横截面

TABLE 13.2 Effects of Incinerator Location on Housing Prices

Dependent Variable: *rprice*			
Independent Variable	**(1)**	**(2)**	**(3)**
constant	82,517.23 (2,726.91)	89,116.54 (2,406.05)	13,807.67 (11,166.59)
y81	18,790.29 (4,050.07)	21,321.04 (3,443.63)	13,928.48 (2,798.75)
nearinc	−18,824.37 (4,875.32)	9,397.94 (4,812.22)	3,780.34 (4,453.42)
y81·nearinc	−11,863.90 (7,456.65)	−21,920.27 (6,359.75)	−14,177.93 (4,987.27)
Other controls	No	*age*, age^2	Full Set
Observations	321	321	321
R-squared	.174	.414	.660

非观测效应或固定效应

利用面板数据，可以把影响因变量的无法观测因素分为两类：一类是恒常不变的，另一类则随时间而变化。我们可将含有单个可观测解释变量的两时期面板数据模型写成

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + \alpha_i + \mu_{it}, t = 1, 2$$

其中α_i一般被称为非观测效应或固定效应，误差μ_{it}常被称为特异性误差或时变误差。

城市犯罪率

$$crmrte_{it} = \beta_0 + \delta_0 d87 + \beta_1 unem_{it} + \alpha_i + \mu_{it}$$

一阶差分估计量

在大多数应用中，收集面板数据的主要理由是为了考虑非观测效应α_i与解释变量相关。

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta \mu_i$$

我们用OLS估计上式，由此得到的估计量称为**一阶差分估计量**(first-differenced estimator)。这里有两个重要的假定，其一是Δx_i和$\Delta \mu_i$无关；其二是Δx_i必须因i的不同而有所变化。后者也很容易理解，由于我们容忍α_i与x_{it}相关，所以我们就不要指望能把α_i对y_{it}的影响与不随时间而变的任何变量的影响分离开来。

6.2 面板数据方法

We use the two years of panel data in SLP75_81.RAW, from Biddle and Hamermesh (1990), to estimate the tradeoff between sleeping and working. In Problem 3 in Chapter 3, we used just the 1975 cross section. The panel data set for 1975 and 1981 has 239 people, which is much smaller than the 1975 cross section that includes over 700 people. An unobserved effects model for total minutes of sleeping per week is

$$slpnap_{it} = \beta_0 + \delta_0 d81_t + \beta_1 totwrk_{it} + \beta_2 educ_{it} + \beta_3 marr_{it} + \beta_4 yngkid_{it} + \beta_5 gdhlth_{it} + a_i + u_{it}, \quad t = 1, 2.$$

$$\widehat{\Delta slpnap} = -92.63 - .227\ \Delta totwrk - .024\ \Delta educ \\ (45.87)\quad (.036)\qquad (48.759) \\ + 104.21\ \Delta marr + 94.67\ \Delta yngkid + 87.58\ \Delta gdhlth \qquad [13.21] \\ (92.86)\qquad (87.65)\qquad (76.60) \\ n = 239,\ R^2 = .150.$$

用两期面板数据做政策分析

令y_{it}为结果变量，并令$prog_{it}$为项目参与虚拟变量，最简单的非观测效应模型为

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 prog_{it} + \alpha_i + \mu_{it}, t = 1, 2$$

如果项目参与仅发生在第二个时期，那么在差分方程中β_1的OLS估计量就有一个非常简单的表达式

$$\hat{\beta}_1 = \bar{\Delta y}_{treat} - \bar{\Delta y}_{control}$$

即我们计算处理组和对照组在这两个时期的平均变化，然后取两者之差便是$\hat{\beta}_1$。

多于两期的差分法

三期面板数据的一阶差分方程

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it1} + ... + \beta_k \Delta x_{itk} + \Delta \mu_{it}, t = 2, 3$$

多于三个时期面板数据的一阶差分方程

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + ... + \alpha_T T_t + \beta_1 \Delta x_{it1} + ... + \beta_k \Delta x_{itk} + \Delta \mu_{it}, t = 2, ..., T$$

如果我们使用多于两期的数据时，欲使用通常的标准误和检验统计量恰当，我们必须假定Δ_{it}是序列无关的。事实上，若假定μ_{it}序列无关且具有恒定方程，则可以证明$\Delta \mu_{i,t}$与$\Delta \mu_{i,t+1}$之间的相关系数为-0.5。若μ_{it}遵循一个稳定的AR(1)模型，则μ_{it}将是序列相关的。只有当μ_{it}遵循一个随机游走时，$\Delta \mu_{it}$才是序列无关的。

固定效应估计法

在某些假定下，起到更好作用的另一种方法是所谓的固定效应变换。考虑仅有一个解释变量的模型，有

$$y_{it} = \beta_1 x_{it} + \alpha_i + \mu_{it}, t = 1, 2, ..., T$$

现在对每个i求方程在时间上的平均，便得到

$$\bar{y}_i = \beta_1 \bar{x}_i + \alpha_i + \bar{\mu}_i$$

用第一个式子减去第二个式子，得到

$$y_{it} - \bar{y}_i = \beta_1 (x_{it} - \bar{x}_i) + (\mu_{it} - \bar{\mu}_i), t = 1, 2, ..., T$$

或

$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it} + \ddot{\mu}_{it}, t = 1, 2, ... T$$

对上述式子进行回归得到的混合OLS估计量被称为固定效应估计量或组内估计量。

6.2 面板数据方法

TABLE 14.1 Fixed Effects Estimation of the Scrap Rate Equation

Dependent Variable: log(*scrap*)

Independent Variables	Coefficient (Standard Error)
d88	−.080 (.109)
d89	−.247 (.133)
grant	−.252 (.151)
$grant_{-1}$	−.422 (.210)
Observations	162
Degrees of freedom	104
R-squared	.201

虚拟变量回归

对每个i估计一个截距的方法，是连同解释变量一起，给每一个横截面观测单位安排一个虚拟变量，这一方法被称为虚拟变量回归。

计算$\hat{\alpha}_i$

在做了固定效应之后，要计算$\hat{\alpha}_i$相当容易

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - ... - \hat{\beta}_k \bar{x}_{ik}, i = 1, 2, ..., N$$

总截距

在固定效应(FE)估计中所报告的截距是$\hat{\alpha}_i$在i上的平均值。

固定效应还是一阶差分

在$T = 2$时，FE和FD的估计值及其全部检验统计量都完全一样。

固定效应还是一阶差分

在$T \geq 3$时，在FE假定下，两者都是无偏且一致的。对较大的N和较小的T，FE和FD之间的选择关键在其估计量的相对效率。当μ_{it}无序列相关时，固定效应法比一阶差分更有效。当μ_{it}遵循随机游走，那么一阶差分法更好。在许多情形下，μ_{it}表现出某种正的序列相关，却未必达到一个随机游走的程度，这时要比较FE和FD估计量的效率就不那么容易了。

固定效应还是一阶差分

将这两种方法都试下常常是一个好主意：如果结果都差不多，也就无所谓了。

随机效应模型

假定一个非观测效应模型

$$y_{it} = \beta_0 + \beta_1 x_{it1} + ... + \beta_k x_{itk} + \alpha_i + \mu_{it}$$

如果假定非观测效应α_i与每个解释变量都无关

$$Cov(x_{itj}, \alpha_i) = 0, t = 1, 2, .., T; j = 1, 2, ..., k$$

上述方程就成为一个随机效应模型。定义复合误差项$\nu_{it} = \alpha_i + \mu_{it}$，则

$$Corr(\nu_{it}, \nu_{is}) = \sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\mu^2), t \neq s$$

随机效应模型

可以通过GLS变换以消去误差中的序列相关，定义

$$\theta = 1 - [\sigma_\mu^2/(\sigma_\mu^2 + T\sigma_\alpha^2)]^{1/2}$$

它介于0与1之间。变换后的方程是

$$y_{it} - \theta\bar{y}_i = \beta_0(1-\theta) + \beta_1(x_{it1} - \theta\bar{x}_{i1}) + ... + \beta_k(x_{itk} - \theta\bar{x}_{ik}) + (\nu_{it} - \theta\bar{\nu}_i)$$

GLS估计量就是上述方程的混合OLG估计量，用$\hat{\theta}$代替θ的可行GLS估计量被称为随机效应估计量。此外，我们可以看到，当$\theta = 0$时，得到混合OLS，而当$\theta = 1$时则得到FE。

准除均值误差

$$\nu_{it} - \theta\bar{\nu}_i = (1-\theta)\alpha_i - \mu_{it} - \theta\bar{\mu}_i$$

6.2 面板数据方法

TABLE 14.2 Three Different Estimators of a Wage Equation

Dependent Variable: log(*wage*)

Independent Variables	Pooled OLS	Random Effects	Fixed Effects
educ	.091 (.005)	.092 (.011)	———
black	−.139 (.024)	−.139 (.048)	———
hispan	.016 (.021)	.022 (.043)	———
exper	.067 (.014)	.106 (.015)	———
$exper^2$	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

随机效应还是固定效应

由于固定效应容许非观测效应与解释变量相关，而随机效应则不然，普遍认为FE是更令人信服的工具。随机效应在某些特定情形中仍可适用。最明显的是，若关键解释变量不随着时间而变化，我们就不能用FE估计其对y的影响。

豪斯曼检验

相当常见的是，研究者同时适用随机效应和固定效应，然后规范地检验时变解释变量系数的统计显著差别。

相关随机效应方法

在某些应用中，我们可以把α_i合理地当作随机变量应用。假设α_i与x_{it}的平均水平相关，则

$$\alpha_i = \alpha + \gamma \bar{x}_i + r_i$$

其中，假设r_i与每个x_{it}都不相关，即$Cov(\bar{x}_i, r_i) = 0$。 运用相关随机效应(CRE)方法，得到

$$y_{it} = \beta x_{it} + \alpha + \gamma \bar{x}_i + r_i + \mu_{it} = \alpha + \beta x_{it} + \gamma \bar{x}_i + r_i + \mu_{it}$$

对上式进行RE回归分析，得到CRE估计值。可得到$\hat{\beta}_{CRE} = \hat{\beta}_{FE}$。考虑CRE的理由起码有两个，其一，CRE方法提供了一个简单、正式地选择FE和RE的方法；其二是它提供了在固定效应分析中包含不随时间变化解释变量的一个途径，例如$y_{it} = \alpha + \beta x_{it} + \gamma \bar{x}_i + \delta z_i + r_i + \mu_{it}$。

把面板数据方法用于其他的数据结构

各种面板数据方法都可以用于一些不涉及时间的数据结构。

As an example, Geronimus and Korenman (1992) used pairs of sisters to study the effects of teen childbearing on future economic outcomes. When the outcome is income relative to needs—something that depends on the number of children—the model is

$$\log(incneeds_{fs}) = \beta_0 + \delta_0 sister2_s + \beta_1 teenbrth_{fs} + \beta_2 age_{fs} + other\ factors + a_f + u_{fs}, \quad [14.18]$$