第六章 面板数据的多元回归分析

- 6.1 引言
- 6.2 联合横截面数据的回归分析
- 6.3 面板数据回归模型
- 6.4 随机效应模型
- 6.5 随机效应模型和固定效应模型的选择

1

- 6．1 引言
- 对相同的单元在不同的时间点得到多个样本
- 在横截面和时间序列两个维度上都有数据
- 数据显然容易违背独立观测样本的要求
- 数据可能会导致观测值之间存在相关性
- 不可观测因素可能对某些单元有固定的影响
- 在各时间点都存在的对一个单元的共同因素的影响效果称为固定效应

2

6．2 联合横截面数据的回归分析

6．2．1 联合横截面数据

- 联合横截面数据可以很快地增大样本量
- 两个或两个以上时间点的样本,具有随机选取的特征，可以直接使用横截面的分析方法
- 直接把联合横截面数据看成横截面进行回归分析
- 两个或多个时间点的数据之间可能存在水平的漂移,加入时间虚拟变量来控制
- 通过年份虚拟变量与某些变量的乘积项来考察该变量在不同年份的影响

3

- 例教育水平在不同年份，性别在不同时代的差别

$$\log(wage) = \beta_0 + \delta_0 y85 + \beta_1 educ + \delta_1 y85 \cdot educ + \beta_2 \exp er + \beta_3 \exp er^2 + \beta_4 union + \beta_5 female + \delta_5 y85 \cdot femal + \varepsilon$$

- δ_1 刻画了7年后，每年的教育收益的改变量
- 7年间男女收入差异变化的零假设为 $H_0 : \delta_5 = 0$
- 得到估计模型为

$$\log(wage) = 0.459 + 0.118 y85 + 0.0747 educ + 0.0185 y85 \cdot educ + 0.0296 \exp er$$

- **(0.093) (0.124) (0.0067) (0.0094) (0.0036)**

$$-0.0004 \exp er^2 + 0.0202 union - 0.317 female + 0.085 y85 \cdot \exp er$$

- **(0.00008) (0.030) (0.037) (0.051)**

4

- 1978年的教育收益是7.5%，1985年的教育收益是9.35%，比1978年高了1.85%；两个时期差异的t统计量是1.85/0.94=1.97
- 两个时期性别差异降低了8.5%，t统计量是8.5/5.1=1.67，在5%的单边检验是显著的
- 对模型中的每一个系数都使用与年虚拟变量的乘积，就相当于分别对两年的数据进行回归
- 不同时间点数据的差异不仅是水平，还可能有方差的不同,要使用更复杂的方法

5

6．2．2 结构改变的Chow检验

- 不同的时段横截面是否一致可通过F检验来考察
- 有两个时期的数据
- 限制模型为对所有样本回归得到残差平方和SSR
- 完全模型分别对两个时期数据进行回归，分别得到残差平方和SSR1和SSR2
- 也可对其中一年样本使用年虚拟变量乘积项，对年虚拟变量和交叉乘积项进行联合显著性检验
- 多时期的数据，同样考虑具有不同的截距项，也可考虑其中的一些变量具有不同的斜率项系数

6

- 除基准年外，其他年变量都要与其中一个，几个或全部变量进行乘积加入模型，会显得比较繁琐
- T个时期的数据
- 限制模型对所有样本进行回归残差平方和为SSR
- 完全模型分别对T个时期数据进行横截面回归，残差平方和 $SSR_u = SSR_1 + SSR_2 + \cdots + SSR_T$
- 检验统计量为 $F = \frac{n - T(k+1)}{(T-1)k} \cdot \frac{SSR - SSR_u}{SSR}$ $n = n_1 + n_2 + \cdots + n_T$
- 只对模型系数在不同时期间存在差异进行检验
- 不同时期间存在异方差时，检验不稳健，功效低

- 6．2．3 联合横截面数据的策略分析

通过事件发生前后的数据来分析事件或策略的影响

- 考虑一个垃圾处理站建立对1999年和2002年房价

$$rprice = 8.252 - 1.882 nearl$$

1999年：（0.2654） （0.5828）

$$rprice = 10.131 - 3.069 nearl$$

2002年： （0.3093） （0.5828）

结果显示在垃圾处理站建立之前和之后两个地区的房价差异都是非常显著的

要回答的是建垃圾处理站是否影响了这一地区房价

- 两个时期，两个地区房价的差异是：
- -3.069-(-1.882)=-1.187
- 这一差异是否显著，还需要它的标准差
- 通过回归模型

$$rprice = \beta_0 + \delta_0 y02 + \beta_1 nearl + \delta_1 y02 \cdot nearl + \varepsilon$$

估计的结果，第一列表示只使用虚拟变量的模型

第二列的模型控制了房子的年限，

第三列控制了影响房价的其他变量，与中心城区距离，站地面积，建筑面积，居室数，卫生间数等

自变量	模型1	模型2	模型3	模型4	模型5
截距项	8.252 (0.273)	8.912 (0.241)	1.381 (1.117)	11.29 (0.31)	0.765 (0.416)
y02	1.879 (0.405)	2.132 (0.344)	1.393 (0.280)	0.193 (0.248)	0.162 (0.028)
nearl	-1.882 (0.488)	0.940 (0.481)	0.378 (0.445)	-0.340 (0.055)	0.032 (0.047)
y02nearl	-1.186 (0.746)	-2.192 (0.636)	-1.418 (0.499)	-0.063 (0.083)	-0.132 (0.052)
控制变量	No	Age，^2	所有变量	No	所有变量
样本量 R方	321 0.174	321 0.414	321 0.660	321 0.239	321 0.724

- 房价采用对数形式，得到模型4和模型5
- 这一方法可以对发生事件或政策、策略的实施的影响进行分析，也被称为自然试验
- 需要有两个样本组
- 分析组，即受事件影响的那些样本
- 控制组，是不受事件影响的那些样本
- 回归模型可以表示为

$$y = \beta_0 + \delta_0 D2 + \beta_1 A + \delta_1 D2 \cdot A + otherfactors + \varepsilon$$

- 乘积项系数表示组间差异的差异

$$\delta_1 = (\bar{y}_{2,A} - \bar{y}_{2,B}) - (\bar{y}_{1,A} - \bar{y}_{1,B})$$

- 6．3 面板数据回归模型
- 6．3．1 面板数据的模型特点
- 面板数据需要通过加入虚拟变量的方法而使用联合横截面的回归方法来给出所需要的结果
- 年份虚拟变量的回归模型为

$$y_{it} = \delta_1 + \delta_2 d2_t + \cdots + \delta_T dT_t + \beta_1 x_{1it} + \cdots + \beta_k x_{kit} + \alpha_i + \varepsilon_{it}$$

- 模型中的α_i 是面板数据中的一个重要特点
- 称为不可观测因素或者称为是固定效应
- 干扰项 ε_{it} 中所包含的也是不可观测的因素，但这一项中的不可观测因素是随时期会有变化
- 而 α_i 中所包含的是在所有时期不会变化的部分

- 两个问题影响模型系数的一致估计
- 假设固定效应与解释变量不相关
- 不同单元的截距相同
- 直接联合横截面回归，相当于把干扰项假设为

$$v_{it} = \alpha_i + \varepsilon_{it}$$

- 称为复合干扰项
- 若假定原来干扰项间彼此无关 $Cov(\varepsilon_{it}, \varepsilon_{js}) = 0$
- 对同一单元复合干扰项则有 $Cov(v_{it}, v_{is}) = Var(\alpha_i)$
- 干扰项在时间方向是序列相关，除非 $Var(\alpha_i) = 0$

13

- 6．3．2 面板数据分析的差分模型
- 把同一单元的数据在不同时期之间进行差分而消除截距项
- 两个时期的模型

$$y_{it} = \beta_0 + \delta_1 d_1 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it}$$

- 两个时期的方程相减得到

$$y_{i2} - y_{i1} = \delta_1 + \beta_1(x_{i21} - x_{i11}) + \cdots + \beta_k(x_{i2k} - x_{i1k}) + \varepsilon_{i2} - \varepsilon_{i1}$$

$$\Delta y_i = \delta_1 + \beta_1 \Delta x_{1i} + \cdots + \beta_k \Delta x_{ki} + \Delta\varepsilon_i$$

- 差分模型排除了解释变量与因变量的后滞项有关
- 注意解释变量在不同的时期要有变化

14

- 单元剔除，要注意选样偏差的问题
- 差分很可能会使自变量的变化幅度大幅减小
- 多时期模型需要使用年份虚拟变量

$$y_{it} = \delta_1 + \delta_2 d2_t + \cdots + \delta_T dT_t + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it}$$

- 对相邻的两个时期的模型对应相减得到

$$\Delta y_{it} = \delta_2(d2_t - d2_{t-1}) + \cdots + \delta_T(dT_t - dT_{t-1}) + \beta_1(x_{it1} - x_{it-11}) + \cdots + \beta_k(x_{itk} - x_{it-11}) + \varepsilon_{it} - \varepsilon_{it-1}$$

$$\Delta y_{it} = \delta_2 \Delta d2_t + \cdots + \delta_T \Delta dT_t + \beta_1 \Delta x_{it1} + \cdots + \beta_k \Delta x_{itk} + \Delta\varepsilon_{it}$$

- 这一模型的估计需要假设 $Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = 0$

15

- 若原模型的假设是 $Cov(\varepsilon_{it}, \varepsilon_{it-1}) = 0 \quad Var(\varepsilon_{it}) = \sigma^2$
- 则差分模型干扰项将存在相关性

$$Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = Cov(\varepsilon_{it} - \varepsilon_{it-1}, \varepsilon_{it-1} - \varepsilon_{it-2})$$

$$= Cov(\varepsilon_{it}, \varepsilon_{it-1} - \varepsilon_{it-2}) - Cov(\varepsilon_{it-1}, \varepsilon_{it-1}) + Cov(\varepsilon_{it-1}, \varepsilon_{it-2})$$

$$= 0 - Var(\varepsilon_{it-1}, \varepsilon_{it-1}) + 0 = -\sigma^2$$

- 自相关系数为

$$\rho(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = \frac{Cov(\varepsilon_{it} - \varepsilon_{it-1}, \varepsilon_{it-1} - \varepsilon_{it-2})}{\sqrt{Var(\varepsilon_{it} - \varepsilon_{it-1}) * Var(\varepsilon_{it-1} - \varepsilon_{it-2})}}$$

$$= -\sigma^2 / \sqrt{2\sigma^2 \times 2\sigma^2} = -0.5$$

16

- 若干扰项是一平稳的AR（1）序列 $\varepsilon_{it} = \rho\varepsilon_{it-1} + \eta_{it}$

$$Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = Cov(\varepsilon_{it}, \varepsilon_{it-1}) - Cov(\varepsilon_{it}, \varepsilon_{it-2}) - Cov(\varepsilon_{it-1}, \varepsilon_{it-1}) + Cov(\varepsilon_{it-1}, \varepsilon_{it-2})$$

$$= \frac{\rho}{1-\rho^2}\sigma_\eta^2 - \frac{\rho^2}{1-\rho^2}\sigma_\eta^2 - \frac{1}{1-\rho^2}\sigma_\eta^2 + \frac{\rho}{1-\rho^2}\sigma_\eta^2 = \frac{\rho-1}{1+\rho}\sigma_\eta^2$$

$$\rho(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = (\frac{\rho-1}{1-\rho}\sigma_\eta^2) / \sqrt{\frac{2}{1+\rho}\sigma_\eta^2 \times \frac{2}{1+\rho}\sigma_\eta^2} = \frac{\rho-1}{2}$$

- 只有当干扰项是随机游动序列时，才有

$$Cov(\Delta\varepsilon_{it}, \Delta\varepsilon_{it-1}) = 0$$

因此使用差分模型时，需检验差分序列的自相关性

通过联合横截面回归得到模型(6.6)的残差序列

$$\Delta\varepsilon_{it} = \xi_{it}$$

17

- 得到向量自回归模型

$$\xi_{it} = \rho\xi_{it-1} + e_{it} \qquad t = 3, \cdots T, i = 1, \cdots, n$$

- 通过t检验可以检验零假设 $H_0 : \rho = 0$
- 若检验结果拒绝零假设，需采用可行广义最小二乘方法（FGLS）来纠正干扰项存在自相关的影响
- 可使用异方差稳健的方法给出估计的标准差
- 例：用1981-1987年数据来给出北卡各县的犯罪率

18

$$\Delta\log(crmr_{it}) = 0.008 - 0.100d83 - 0.048d84 - 0.005d85$$
$$(0.017)\quad (0.024)\quad (0.024)\quad (0.023)$$
$$[0.014]\quad [0.022]\quad [0.020]\quad [0.025]$$
$$+ 0.028d86 + 0.041d87 + 0.398\Delta\log(polpc) - 0.238\Delta\log(prbcom)$$
$$(0.024)\quad (0.024)\quad (0.027)\quad (0.018)$$
$$[0.021]\quad [0.024]\quad [0.101]\quad [0.039]$$
$$- 0.165\Delta\log(prbpris) - 0.022\Delta\log(avgsem) - 0.327\Delta\log(prbarr)$$
$$(0.026)\quad (0.022)\quad (0.030)$$
$$[0.045]\quad [0.025]\quad [0.056]$$
$$n = 540, R^2 = 0.433, \overline{R}^2 = 0.422$$

19

- 提高抓捕和定罪的概率，提高服刑的年限能显著的抑制一个地区的犯罪率
- 人均警察的数量与预期相反，一个地区的人均警察越多，反而犯罪率越高（这是犯罪率研究中的一个普遍现象）
- 这一现象可能是受内生因果关系的影响
- 犯罪率是实际报告的犯罪率，一个地区警察越多，报告和抓捕的犯罪可能更多

20

6．3．3 固定效应模型

- 从没有年虚拟变量的基本模型出发

$$y_{it} = \beta_0 + \beta_1 x_{it1} + \cdots + \beta_k x_{itk} + \alpha_i + \varepsilon_{it}$$

- 把所有关于某单元的方程进行平均得

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{i1} + \cdots + \beta_k \bar{x}_{ik} + \alpha_i + \bar{\varepsilon}_i$$

- 用平均后的模型减原来的模型得

$$y_{it} - \bar{y}_i = \beta_1(x_{it1} - \bar{x}_{i1}) + \cdots + \beta_k(x_{itk} - \bar{x}_{ik}) + \varepsilon_{it} - \bar{\varepsilon}_i$$

$$\tilde{y}_{it} = \beta_1 \tilde{x}_{it1} + \cdots + \beta_k \tilde{x}_{itk} + \tilde{\varepsilon}_{it}$$

- 自变量和干扰项也进行了相同的变换，这一变换称为固定效应转换

21

- 在干扰项与解释变量不相关，在不同时期之间是互不相关的假设下，得到的估计是无偏
- 这类模型中不能包含对某些样本单元在各样本时期都不变的量
- 使用OLS估计固定效应模型需要假设干扰项是同方差的，且在不同时期是序列不相关的
- 要注意自由度的变化，模型没有截距项
- 模型的样本仍然是NT个，但自由度为N(T-1)-k个
- 各时期为常数的变量可通过与其他变量的乘积项而进入模型

22

6．3．4 虚拟变量回归模型

- 样本单元数量不大时，可对每一个单元设一个虚拟变量，回答各单元的固定效应
- 同样可以给出使用剔除时间均值的固定效应模型估计相同的结果，相应的标准差和检验统计量
- 没有了模型自由度的问题，R方通常比较高
- 固定效应模型给出的结果，通过平均水平的模型可得到关于某些单元是否有显著的固定效应

$$\hat{\alpha}_i = \bar{y}_i - \hat{\beta}_1 \bar{x}_{i1} - \cdots - \hat{\beta}_k \bar{x}_{ik}$$

23

6．3．5 固定效应和差分模型的选择

- 只有两个时期的时候，固定效应模型和差分模型实际上是一样的，使用差分模型更简便
- 不能以是否无偏来选择模型
- N比较大，而T相对比较小时，选择模型的标准就要考虑它们的相对有效性
- 干扰项不相关时固定效应模型比差分模型更有效
- 干扰项序列具有比较高的正相关，差分模型要优于固定效应模型
- 最常见的是干扰项序列存在一定程度的正相关

24

要比较固定效应模型和差分模型谁更有效比较困难

- 固定效应模型不能直接检验残差序列自相关性
- 对差分模型的残差进行检验，如果该序列是不相关的，则使用差分模型更有效，如果该序列存在比较显著的负相关，可能固定效应模型更合适
- T很大，横截面单元数量不大，固定效应的模型的结果变得非常敏感
- 当使用的数据沿时间方向是单位根过程时，固定效应可能出现伪回归的问题

固定效应和差分模型给出结果差异较大，很难选择

25

6．3．6 对不完整面板数据的处理

- 一些单元缺少个别时期数据称为不完整面板数据
- 单元i有T_i个时期的数据，只用这T_i个数据取均值

总的样本观测值为$T_1+T_2+\cdots+T_N$

每个单元减少了一个自由度

其余类似于完全面板数据

如果单元数据的缺失与不可观测因素有关，则模型给出的估计可能是有偏的

26

6．4 随机效应模型

- 考虑带有不可观测因素的模型

$$y_{it}=\delta_1+\delta_2 d2_t+\cdots+\delta_T dT_t+\beta_1 x_{1it}+\cdots+\beta_k x_{kit}+\alpha_i+\varepsilon_{it}$$

固定效应或差分模型中需要消去不可观测效应α_i是因为它们与解释变量存在相关性

假定不可观测效应在任何时期都与解释变量不相关

$$Cov(x_{ij},\alpha_i)=0, t=1,2,\cdots T,\ j=1,\cdots,k$$

满足这一假设条件加上固定效应模型其他假设的模型称为随机效应模型

27

随机效应模型假设成立时，只用一个时期的横截面数据就可给出模型系数的一致估计，用联合横截面回归方法包含各时期的虚拟变量使用所有数据

忽略了随机效应模型的一个重要特征

模型的复合干扰项为 $\nu_{it}=\alpha_i+\varepsilon_{it}$

联合横截面模型中

$$y_{it}=\delta_1+\delta_2 d2_t+\cdots+\delta_T dT_t+\beta_1 x_{it1}+\cdots+\beta_k x_{itk}+\nu_{it}$$

模型的干扰项序列在各时期之间是彼此相关的

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

28

OLS估计的联合横截面模型忽略了干扰项序列之间的相关性，需用广义OLS方法来给出具有干扰项序列自相关模型的估计

定义调整系数 $\lambda=1-[\sigma_\varepsilon^2/(\sigma_\varepsilon^2+T\sigma_\alpha^2)]^{1/2}$

用均值模型乘以调整系数再减回原来的模型得

$$y_{it}-\lambda\bar{y}_i=\delta_1(1-\lambda)+\delta_2(1d2_t-\lambda/T)+\cdots+\delta_T(1dT_t-\lambda/T)+$$
$$+\beta_1(x_{it1}-\lambda\bar{x}_{i1})+\cdots+\beta_k(x_{kit}-\lambda\bar{x}_{ik})+\nu_{it}-\lambda\bar{\nu}_1$$

调整系数介于0和1之间，是去除部分均值的变换模型

变换后的模型比固定效应和差分模型的优点是它可以包含在各时期有相同变化率的变量

29

调整系数的估计

用联合横截面或固定效应的OLS，估计 $\hat{\alpha}_i$ 和残差 $\hat{\varepsilon}_{it}$

方差的估计分别为 $\hat{\sigma}_\alpha^2=Var(\alpha_i)=\frac{1}{n-1}\sum_{i=1}^n\hat{a}_i^2$

$$\hat{\sigma}_\varepsilon^2=Var(\varepsilon_{it})=\frac{1}{T-1}\sum_{t=1}^T\hat{\varepsilon}_{it}^2$$

$$\hat{\lambda}=1-[\hat{\sigma}_\varepsilon^2/(\hat{\sigma}_\varepsilon^2+T\hat{\sigma}_\alpha^2)]^{1/2}$$

也可从联合横截面OLS估计的残差序列得到方差估计

$$\hat{\sigma}_\alpha^2=\frac{1}{\frac{1}{2}nT(T-1)-k+1}\sum_{i=1}^n\sum_{t=1}^{T-1}\sum_{s=t+1}^T\hat{\nu}_{it}\hat{\nu}_{is}$$

$$\hat{\sigma}_\varepsilon^2=\hat{\sigma}_\nu^2-\hat{\sigma}_\alpha^2$$

30

固定T时，随着N增大，可行GLS是无偏的且渐进正态

N很小而T又很大时，随机效应可行GLS估计性质不好

固定效应，联合横截面和随机效应模型之间的关系

当 $\lambda = 0$ 时，随机效应模型变为联合横截面模型

当 $\lambda = 1$ 时，随机效应模型等价于固定效应模型

(6.18)中的干扰项进行分解得

$$v_{it} - \lambda \bar{v}_1 = (1-\lambda)\alpha_i + \varepsilon_{it} - \lambda \bar{\varepsilon}_i$$

随机效应模型中由于不可观测效应与解释变量的相关性而导致随机效应模型不一致的程度为 $(1-\lambda)$

联合横截面，固定效应和随机效应三种模型估计差异

解释变量	教育年限	黑色人种	西班牙裔	工作年限	工作年限2	婚姻状况	协会
联合横截面	0.091 (0.005)	-0.139 (0.024)	0.016 (0.021)	0.067 (0.014)	-0.0024 (0.0008)	0.108 (0.016)	0.182 (0.017)
随机效应	0.092 (0.011)	-0.139 (0.048)	0.022 (0.043)	0.106 (0.015)	-0.0047 (0.0007)	0.064 (0.017)	0.106 (0.018)
固定效应					-0.0052 (0.0007)	0.047 (0.018)	0.080 (0.019)

6．5　随机效应模型和固定效应模型的选择

- 当样本来自比较大的群体，不可观测效应是一个需要估计的参数，因此多采用固定效应模型
- 确定采用随机效应模型的时，需要注意不可观测的效应是否会与解释变量之间存在相关性
- 比较固定效应模型和随机效应模型实际上是对不可观测效应与解释变量之间是否存在相关性的一个检验
- 随机效应是否合适的Hausman检验统计量

总结：

- 介绍了联合横截面数据和面板数据的特点
- 讨论使用联合横截面数据的估计方法
- 是否可以使用联合横截面数据的Chow检验
- 消除固定效应影响的差分模型和固定效应模型
- 控制固定效应的虚拟变量模型；
- 根据干扰项序列相关性对差分和固定效应进行选择
- 固定效应与解释变量不相关时的随机效应模型
- 随机效应调整系数的估计方法
- 固定效应和随机效益模型的选择标准

再见！