政治经济学前沿方法论与量化分析

第八讲 面板回归

上课地点： 善斋306C
上课时间：周二第六大节

龙治铭
善斋307C
zhiminglong@tsinghua.edu.cn

清華大學
Tsinghua University

目录

CONTENTS

1

面板回归的理论基础

面板数据（Panel Regression）

※前面提到了三种数据类型：
横截面数据：同一随机变量同一时空的不同观测值，如学生期末考试成绩
时间序列：同一随机变量同一空间，不同时间的观测值，如中国年度GDP
面板数据：二者的结合两个维度，不同时间，不同空间的观测值：各省各年末人口数

※空间自相关(spatial autocorrelation)：
空间自相关不一定局限于“地理空间”的意义，而是指在时间为t的横截面数据中，随机扰动项之间存在相关性：$Cov(\epsilon_{i,t}, \epsilon_{j,t}) \neq 0$
(OLS理想条件不成立)
例：在限制了优秀率的情况下，期末考试成绩不仅仅取决于自身努力程度，也取决于相对努力程度。

FIRM ID	TOTAL SALES	VALUE ADDED	Error
firm 1	X_{1t}	Y_{1t}	ϵ_{1t}
firm 2	X_{2t}	Y_{2t}	ϵ_{2t}
firm 3	X_{3t}	Y_{3t}	ϵ_{3t}
⋮			
firm N	X_{Nt}	Y_{Nt}	ϵ_{Nt}

※序列自相关（serial autocorrelation)：
对于某个具体的空间ID，时序的观测值很有可能具有相关性（当时间较短的时候尤其如此）

※空间自相关和序列自相关的类型决定了如何为面板数据选择合适的模型

YEAR ID	TS of firm i	VA of firm i	Error
year 1	X_{i1}	Y_{i1}	ϵ_{i1}
year 2	X_{i2}	Y_{i2}	ϵ_{i2}
year 3	X_{i3}	Y_{i3}	ϵ_{i3}
⋮			
year T	X_{iT}	Y_{iT}	ϵ_{iT}

面板数据的表示和优缺点

※面板数据是Panel data的直译，比较生硬，在有的统计学教材中被称为：longitudinal data或者crosssectional time-series data

※面板数据的优点：

1）可以考虑个体的异质性：面板数据涉及到个体（个人、公司、省份、国家。。。）的数据，允许控制如文化差异等难以量化和观察的因素以及某些遗漏变量

2）适用于分层抽样

3）增加了更多的信息，降低了多重共线性，增加了观测值个数

4）是内生性问题的一种解决办法：过去可以影响未来，但未来不能影响过去，变量的滞后阶是一个天然的工具变量。以及遗漏变量。

※缺点：数据收集（统计学家对抽样设计的批评），国家之间的相关性等

更多见：Baltagi（2008）, Econometric Analysis of Panel Data

Stock and Watson (2007), Chapter 10: Regression with panel data

country	year	Y	X1	X2	X3
1	2000	6.0	7.8	5.8	1.3
1	2001	4.6	0.6	7.9	7.8
1	2002	9.4	2.1	5.4	1.1
2	2000	9.1	1.3	6.7	4.1
2	2001	8.3	0.9	6.6	5.0
2	2002	0.6	9.8	0.4	7.2
3	2000	9.1	0.2	2.6	6.4
3	2001	4.8	5.9	3.2	6.4
3	2002	9.1	5.2	6.9	2.1

在增长理论里，使用面板数据来检验所谓的经济增长的“收敛性”是常见的文献

面板数据的模型

※面板数据的建模，使用前面的方法要么对每一个个体进行时间序列建模，例如解释世界各国的经济增长，可以对中国、美国各个国家单独进行时间序列建模；或者对于某个时间横截面，使用传统的横截面数据建模（如最小二乘法），例如用股市3000多只股票的市盈率等解释某一天的指数等。

例子：估计航空公司的成本函数

I = airline id, T = year, Q = output in revenue passenger miles, C = total cost in $1; 000, PF = fuel price, LF = load factor

对每个航空公司进行时间序列建模

$$C_t = \beta_1 + \beta_2 Q_t + \beta_3 PF_t + \beta_4 LF_t + u_t$$

几百家航空公司就有几百个模型

对每一年所有的航空公司进行横截面数据建模

$$C_i = \beta_1 + \beta_2 Q_i + \beta_3 PF_i + \beta_4 LF_i + u_i$$

多少年的数据就有多少个模型

※有四种方法可以综合考虑二者：

1）the pooled OLS model
2）least squares dummy variable model (LSDV)
3）fixed effects within group model
4）random effects model

后两种是最为常用的方法

Pooled OLS model

※暴力简单，忽略面板数据的性质，仅仅将所有观测值放在一起，使用OLS：

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

※前提假设：

1）标准的OLS假设

2）个体之间没有任何差异，例子中就是航空公司都一样没有区别，这个假设通常很难符合现实。

※解决办法：

异质性（heterogeneity）或者说个体性（individuality）使得Pooled OLS 的残差值具有（空间）自相关性，第四讲中指出此时OLS不再具有方差最小的性质。两种解决办法（还有其他很多种办法）：

1）引入一个额外的不随时间变化（time-invariant）的解释变量α_i,用来刻画异质性效应（heterogeneity effect）

例如：

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + \alpha_i + u_{it}$$

其中α_i可以是管理水平、所有制形式、老板的性别。。。等等以及所有影响成本的不同因素，但不随时间变化。

发展出固定效应模型 “fixed effects within group model”

2）由于α_i通常是不可观察到的，我们可以假定它是随机的，并包含在随机扰动项内。发展出随机效应模型（ random effects model ）

$$C_{it} = \beta_1 + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + \alpha_i + v_{it}$$

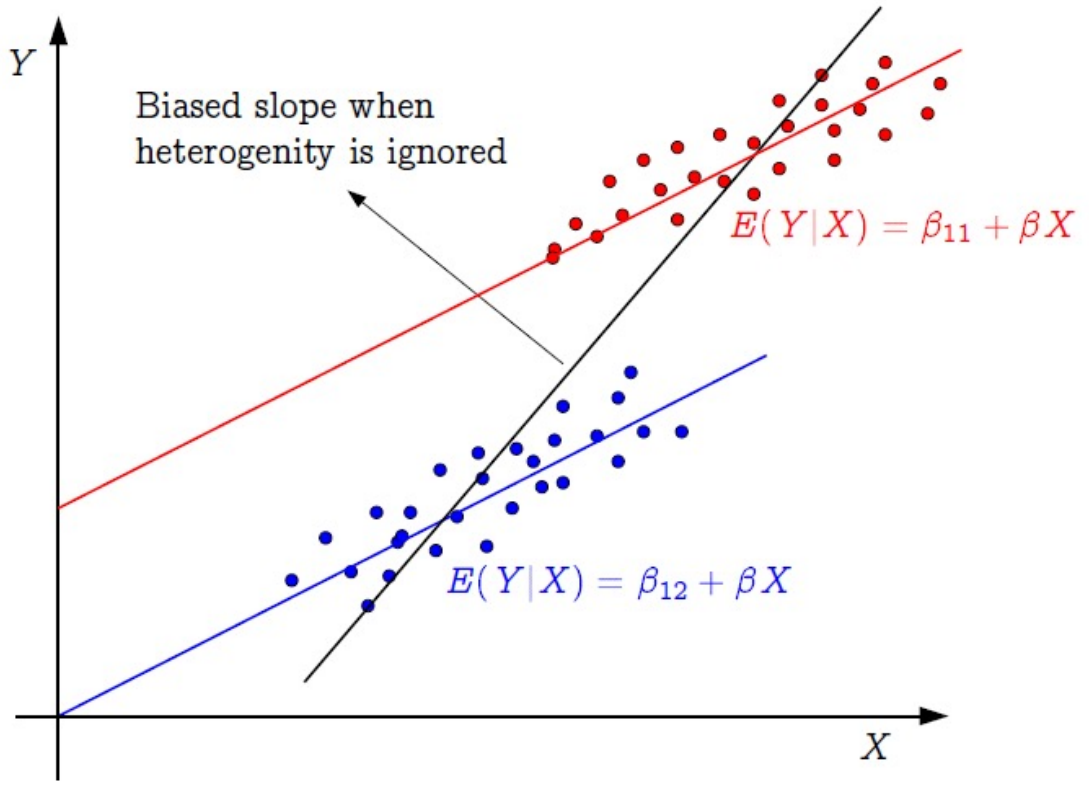
$$v_{it} = \alpha_i + u_{it}$$

The LSDV model

※想法：在上面解决方法1里我们引入了一个不随时间变化的变量α_i,用来刻画异质性效应，既然是time-invariant，那么在时间序列的维度上它是常数，而对于空间上的其他个体，这个特征又是0，显然符合哑变量的特征。

从图中我们可以看见，散点图呈现出如下特征：两组斜率类似但截距不同。如果我们允许不同的组有不同的截距，斜率的有偏估计可以得到避免
LSDV model：（思考一下和Tobit model的异同）

$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$



※缺点：缺点非常显著，当个体数较多的，不得不使用非常多的哑变量，例如，如果有六家航空公司：

$$C_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \alpha_5 D_{5i} + \alpha_6 D_{6i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

过多的哑变量显然会引起共线性问题

※Two-way fixed effects model：允许异质性效应随时间变化

The FE or WITHIN-GROUP model

※固定效应模型（ fixed effects model 也称为within group model ）基于以下简单的想法：既然异质性效应是time-invariant，那么经过差分就可以消去，就避免了引入过多哑变量所导致的共线性问题。

$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it}$$

※方法一mean-corrected model as WG model：

假定均值是：$\bar{C}_i = \beta_{1i} + \beta_2 \bar{Q}_i + \beta_3 \bar{PF}_i + \beta_4 \bar{LF}_i$

相减消去了异质性 $C_{it} - \bar{C}_i = \beta_2(Q_{it} - \bar{Q}_i) + \beta_3(PF_{it} - \bar{PF}_i) + \beta_4(LF_{it} - \bar{LF}_i) + u_{it}$

模型变为

$$c_{it} = \beta_2 q_{it} + \beta_3 pf_{it} + \beta_4 lf_{it} + u_{it}$$

显然，LSDV 和FE models在数学上是等价的，通过以下办法可以还原得到异质性效应：

$$\hat{\beta}_{1i} = \bar{C}_i - \hat{\beta}_2 \bar{Q}_i - \hat{\beta}_3 \bar{PF}_i - \hat{\beta}_4 \bar{LF}_i$$

※方法二:通过差分消去异质性：first-difference FD estimator

$$\Delta C_{it} = \beta_2 \Delta Q_{it} + \beta_3 \Delta PF_{it} + \beta_4 \Delta LF_{it} + \Delta u_{it}$$

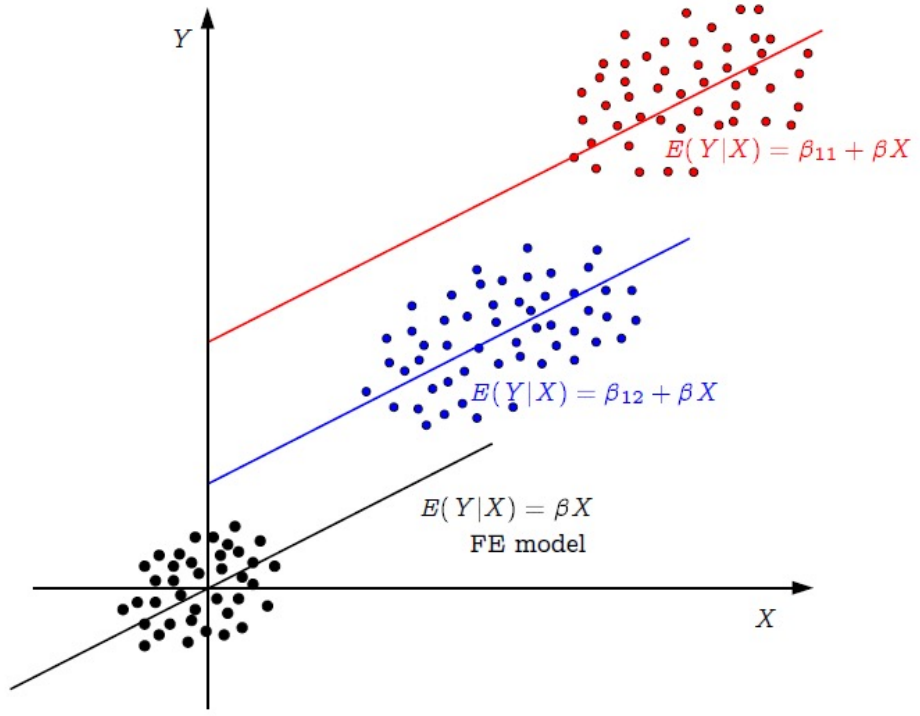
缺陷：

如果随机扰动项不是相关的，那么差分后是自相关的

如果随机扰动项是随机游走，那么差分后是白噪音

※FE还是FD的选择：时间只有两期，二者等价。

时间大于2期，二者均是一致无偏估计量（证明从略），根据随机扰动项性质判断使用哪一个。



FE model的局限性

※FE还是FD的选择：
时间只有两期，二者等价。
时间大于2期，二者均是一致无偏估计量（证明从略），根据随机扰动项性质判断使用哪一个。
随机扰动项无自相关性：FE优于FD
随机扰动项是随机游走：FD优于FE

※局限性：
1）时间很长时，谨慎使用FE，原因很简单，较长时间内异质性可能不是time-invariant
2）可能消除某些已知的time-invariant variable,例如公司位置等
3）长期趋势被消除
4）FE估计量不是efficient estimator (方差不是最小)，但是是一致估计量
5）FE显著地降低了自由度（FE和LSDV本质等价）

The Random Effect RE model

※随机效应模型：在固定效应模型里，我们认为异质性是time-invariant的，但事实上这一假设可能不成立。在随机效应模型里，我们假定截距是一个随机变量：

$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + u_{it} \qquad \beta_{1i} = \beta_1 + \epsilon_i$$

※ RE model的解释：

1）样本中的个体是从一个更大总体中取出来的

2）样本中的个体性具有某种共同的均值β_1

3）个体之间的差异性则体现在ϵ上

4）适用范围：全样本不适用。

例如考察全国各省的经济增长，不适用于RE；考察100家国企的效益，适用于RE。

※表达式和假设：

$$C_{it} = \beta_{1i} + \beta_2 Q_{it} + \beta_3 PF_{it} + \beta_4 LF_{it} + w_{it} \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2) \quad u_{it} \sim N(0, \sigma_u^2)$$

$$w_{it} = \epsilon_i + u_{it} \qquad E(\epsilon_i u_{it}) = 0 \quad E(\epsilon_i \epsilon_j) = 0 \ \ i \neq j$$

$$E(u_{it} u_{is}) = E(u_{it} u_{jt}) = E(u_{it} u_{js}) \ \ i \neq j \ t \neq s$$

w_{it}因此包含两个部分：ϵ_i代表个体误差部分

u_{it}包含个体误差和时间序列误差

※估计方法：可以证明 $COV(w_{it}, w_{is}) = \sigma_\epsilon^2 \neq 0 \quad t \neq s$

因此RE model 的估计不应当使用OLS，最好使用GLS

FE V.S RE

※在固定效应模型和随机效应模型我们应该选择使用哪一个呢？

※ Hausman-Wu test: 第四讲中介绍的HW test可以用来作为判断依据
原因：二者都是一致估计量，比较哪一个更加efficient.
H0: FE和RE估计量没有显著差异
如果H0被拒绝，RE可能不合适，因为ϵ_i很有可能与某个自变量相关。

※经验判断：
1）如果T很大，N很小，RE和FE之间可能没有太大区别
2）如果T很小，N很大， RE和FE之间可能区别很大。如果样本不是从一个更大的总体中随机得到，我们倾向于FE。
3）如果个体误差ϵ_i与某个自变量相关，RE是有偏估计，而FE不是
4）FE不能估计time-invariant变量，但RE可以
5）FE控制所有的time-invariant variables但RE值隐含地控制那些被引入模型中的time-invariant variables

Extension of Panel regression

※面板数据与ARMA模型结合

※面板数据与cointegration结合

※面板数据与logit model结合

※空间计量模型

2

面板回归在Stata中的实现

问题和数据

※面板回归广泛应用于增长理论中“收敛”问题的研究，但事实上早期的文献如巴罗、曼昆、罗默等人在90年代的论文都是有问题的（伪回归）。

※单位根检验：面板数据包含时间维度，同样需要作单位根检验，但如果时间维度较短，可以根据图像和数据性质判断平稳性和趋势项类型（如果使用FE，趋势项或者随机游走可能可以被消除）

我在博士一年级时指出了哈佛大学Torben Iversen 教授使用面板回归时的错误，他们表示感谢：

“The Politics of Opting Out: Explaining Educational Financing and Popular Support for Public Spending” (with Marius Busemeyer). *Socio-Economic Review* 12,April 2014, 299-328.

the University of North Carolina at Chapel Hill in November 2013 and at the Political Economy Seminar at the University of Paris I in December 2013. We thank the participants at these events for many helpful comments and suggestions. We are particularly grateful to Ben Ansell, Julian Garritzmann and Michael

※数据和问题的选择：文献中省级资本存量的估计有严重缺陷，不适合作为回归分析（Long and Herrera,2016,CER），这一工作我尚未完成。Stata自带数据库只有一个变量（卡路里），我在网上找了一个普林斯顿的讲义：

讲义地址： https://www.princeton.edu/~otorres/Panel101.pdf

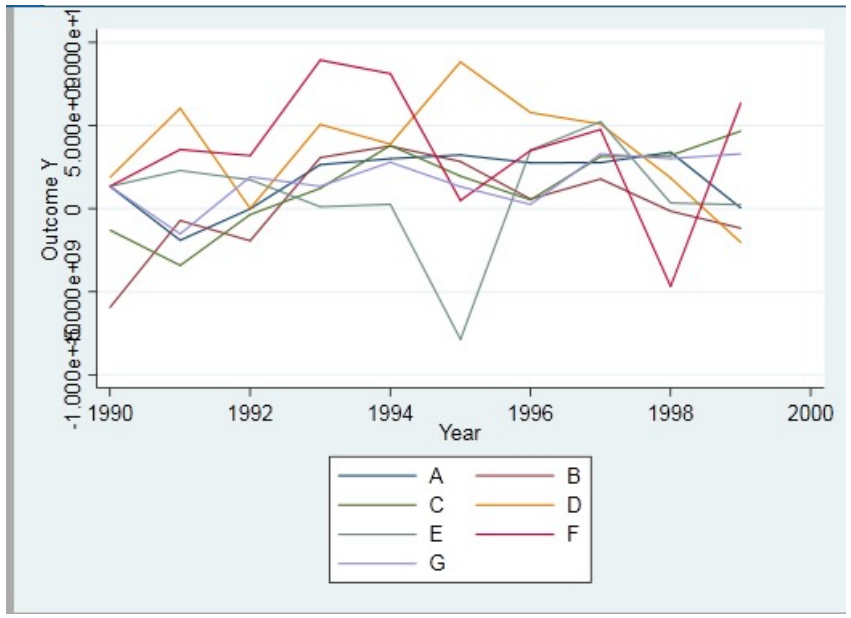
数据地址： https://dss.princeton.edu/training/Panel101.dta

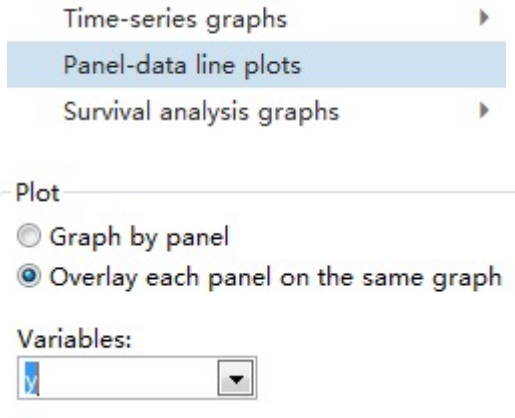
数据和图像

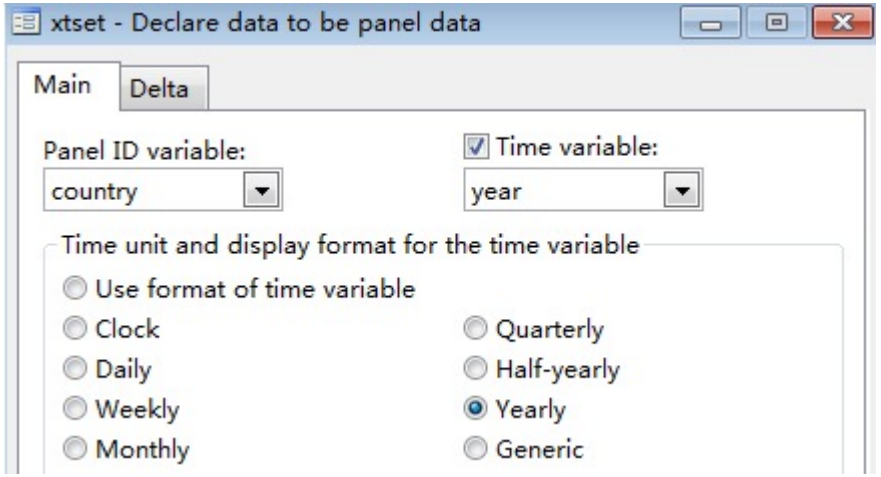
※导入数据：use "C:\Users\Administrator\Desktop\Panel101.dta "

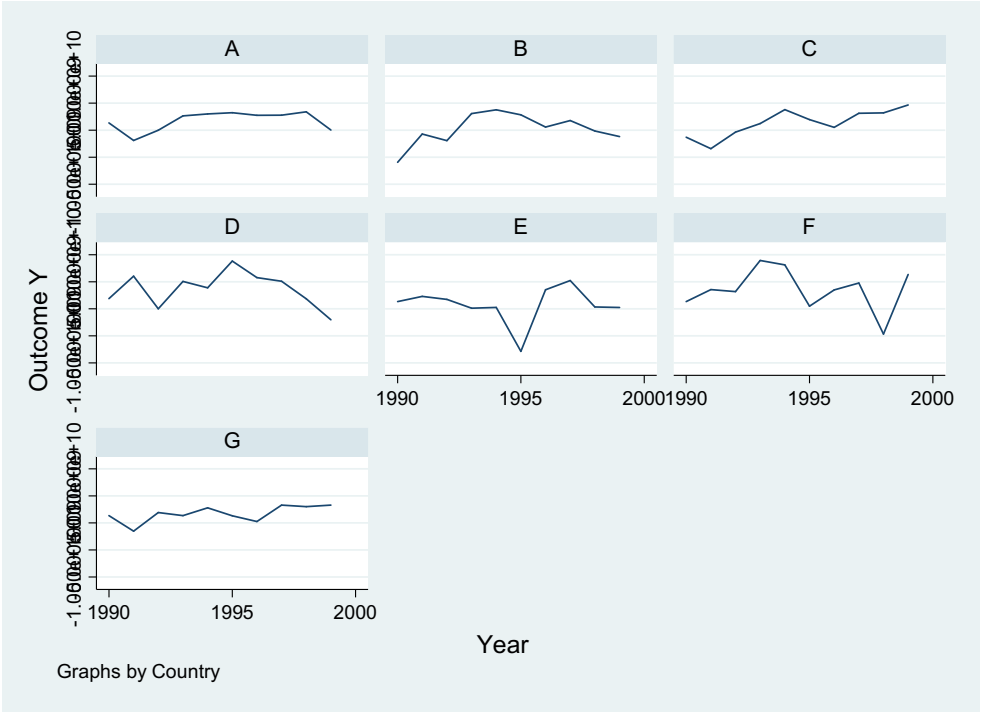
※设定数据为面板数据： xtset country year, yearly

为面板数据绘制图像： xtline y
可以选择是为各个国家分别绘图
还是放在同一坐标系内
xtline y, overlay





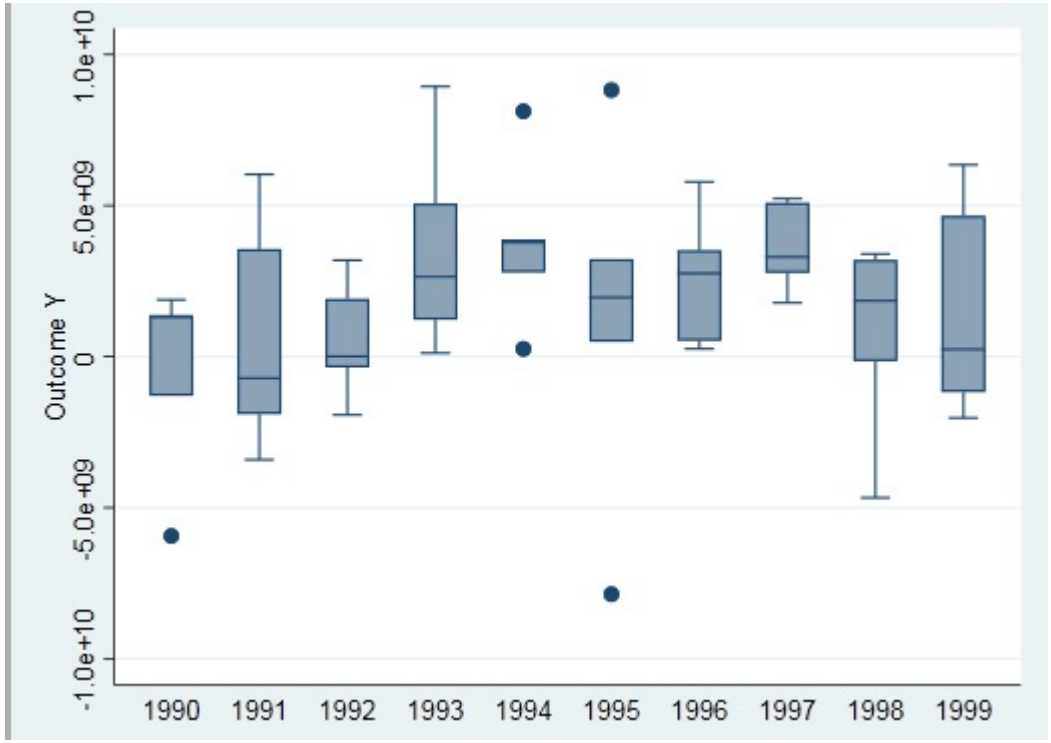


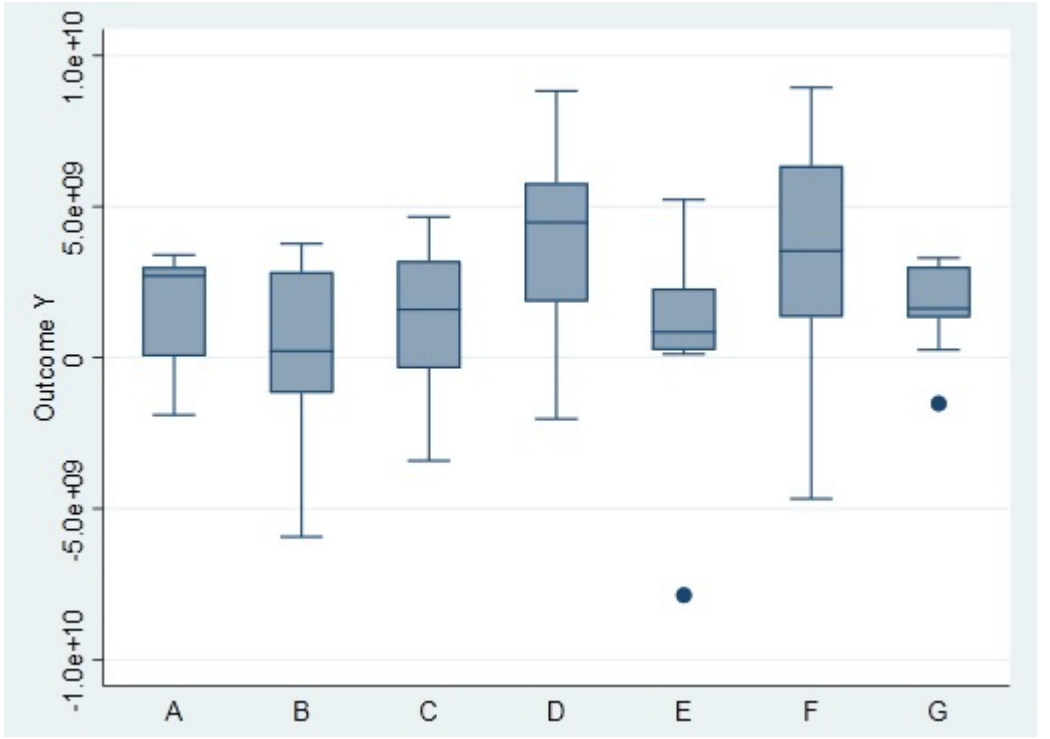


固定效应模型探索

※空间上（国与国之间）是否存在异质性？时间维度上是否存在异质性？
graph box y, over(country)
graph box y, over(year)

箱线图表明，不同国家之间的差异性很大。





LSDV model和FE model 等价

※LSDV model和FE model 是等价的，回归中关于X1的系数也证明了这点

LSDV ：regress y x1 i.country　　FE: xtreg y x1, fe

```
regress y x1 i.country
```

Source	SS	df	MS	Number of obs	=	70
				F(7, 62)	=	2.61
Model	1.4276e+20	7	2.0394e+19	Prob > F	=	0.0199
Residual	4.8454e+20	62	7.8151e+18	R-squared	=	0.2276
				Adj R-squared	=	0.1404
Total	6.2729e+20	69	9.0912e+18	Root MSE	=	2.8e+09

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
country						
B	-1.94e+09	1.26e+09	-1.53	0.130	-4.47e+09	5.89e+08
C	-2.60e+09	1.60e+09	-1.63	0.108	-5.79e+09	5.87e+08
D	2.28e+09	1.26e+09	1.81	0.075	-2.39e+08	4.80e+09
E	-1.48e+09	1.27e+09	-1.17	0.247	-4.02e+09	1.05e+09
F	1.13e+09	1.29e+09	0.88	0.384	-1.45e+09	3.71e+09
G	-1.87e+09	1.50e+09	-1.25	0.218	-4.86e+09	1.13e+09
_cons	8.81e+08	9.62e+08	0.92	0.363	-1.04e+09	2.80e+09

```
Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:                                           Obs per group:
     within  = 0.0747                                         min =        10
     between = 0.0763                                         avg =      10.0
     overall = 0.0059                                         max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_cons	2.41e+08	7.91e+08	0.30	0.762	-1.34e+09	1.82e+09
sigma_u	1.818e+09					
sigma_e	2.796e+09					
rho	.29726926	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(6, 62) = 2.97                     Prob > F = 0.0131
```

※模型的解释： corr(u_i, Xb) = -0.5468 残差值与自变量相关

sigma_u：within groups residual ui的标准差　　sigma_e: residual ei的标准差

Rho：29.72%的方差是因为panel 间差异　三种R2　模型具有整体显著性

LSDV model和FE model 等价

※LSDV model和FE model 是等价的，回归中关于X1的系数也证明了这点

LSDV ：regress y x1 i.country　　FE: xtreg y x1, fe

```
regress y x1 i.country
```

Source	SS	df	MS	Number of obs	=	70
				F(7, 62)	=	2.61
Model	1.4276e+20	7	2.0394e+19	Prob > F	=	0.0199
Residual	4.8454e+20	62	7.8151e+18	R-squared	=	0.2276
				Adj R-squared	=	0.1404
Total	6.2729e+20	69	9.0912e+18	Root MSE	=	2.8e+09

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
country						
B	-1.94e+09	1.26e+09	-1.53	0.130	-4.47e+09	5.89e+08
C	-2.60e+09	1.60e+09	-1.63	0.108	-5.79e+09	5.87e+08
D	2.28e+09	1.26e+09	1.81	0.075	-2.39e+08	4.80e+09
E	-1.48e+09	1.27e+09	-1.17	0.247	-4.02e+09	1.05e+09
F	1.13e+09	1.29e+09	0.88	0.384	-1.45e+09	3.71e+09
G	-1.87e+09	1.50e+09	-1.25	0.218	-4.86e+09	1.13e+09
_cons	8.81e+08	9.62e+08	0.92	0.363	-1.04e+09	2.80e+09

```
Fixed-effects (within) regression               Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:                                           Obs per group:
     within  = 0.0747                                         min =        10
     between = 0.0763                                         avg =      10.0
     overall = 0.0059                                         max =        10

                                                F(1,62)            =      5.00
corr(u_i, Xb)  = -0.5468                        Prob > F           =    0.0289
```

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
x1	2.48e+09	1.11e+09	2.24	0.029	2.63e+08	4.69e+09
_cons	2.41e+08	7.91e+08	0.30	0.762	-1.34e+09	1.82e+09
sigma_u	1.818e+09					
sigma_e	2.796e+09					
rho	.29726926	(fraction of variance due to u_i)				

```
F test that all u_i=0: F(6, 62) = 2.97                     Prob > F = 0.0131
```

※模型的解释： corr(u_i, Xb) = -0.5468 残差值与自变量相关

sigma_u：within groups residual ui的标准差　　sigma_e: residual ei的标准差

Rho：29.72%的方差是因为panel 间差异　三种R2　模型具有整体显著性

RE model

※RE model : xtreg y x1, re
※模型的解释：
corr(u_i, X) = 0 (assumed) 严格的外生型假设
Prob > chi2 = 0.1669 模型不具备整体显著性
※使用HW 检验究竟使用RE还是FE
xtreg y x1, fe
estimates store fixed
xtreg y x1, re
estimates store random
hausman fixed random

```
Random-effects GLS regression                   Number of obs      =        70
Group variable: country                         Number of groups   =         7

R-sq:                                           Obs per group:
     within  = 0.0747                                         min =        10
     between = 0.0763                                         avg =      10.0
     overall = 0.0059                                         max =        10

                                                Wald chi2(1)       =      1.91
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.1669
```

y	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
x1	1.25e+09	9.02e+08	1.38	0.167	-5.21e+08	3.02e+09
_cons	1.04e+09	7.91e+08	1.31	0.190	-5.13e+08	2.59e+09
sigma_u	1.065e+09					
sigma_e	2.796e+09					
rho	.12664193	(fraction of variance due to u_i)				

```
. hausman fixed random
```

	Coefficients (b) fixed	(B) random	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
x1	2.48e+09	1.25e+09	1.23e+09	6.41e+08

```
                b = consistent under Ho and Ha; obtained from xtreg
    B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test:  Ho:  difference in coefficients not systematic

          chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                  =        3.67
        Prob>chi2 =      0.0553
```

P-value=0.0553>0.05，在5%的水平上不能拒绝原假设，使用RE。
在10%的水平上拒绝原假设，使用FE。

Validation和其他检验

※RE model使用GLS，对残差值的要求较为宽松，当然如果是高斯更好
自相关性和异方差性检验方法与第四讲相同，此处从略，或见普林斯顿的讲义

3

参考文献


站在巨人的肩膀上
"If I have seen further it is by standing on the shoulders of Giants. “
by Isaac Newton in 1675

作业

※分组讨论：3人一组，共5组，5月7日介绍你们的研究。每组20分钟

二选一作业：

1）使用1-8讲介绍的方法，任选感兴趣的问题做实证分析（加分作业）

2）实在找不到问题的，在top5期刊找一篇实证分析论文，复刻该论文，介绍该论文的研究思路、理论模型、实证方法、数据处理，主要结论。

谢谢！

马克思主义学院
龙治铭

清華大學
Tsinghua University