面板数据分析方法

汇报人：成思扬

2020年11月26日星期四

面板数据分析方法

- 第五节：固定效应模型估计方法
  - 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
  - 时间固定效应的引入
  - 个体效应残差
- 第六节：面板数据分析实例
- 第七节：面板数据运用常见问题

第五节 固定效应模型估计方法

- $Y_{it} = X_{it}^T\beta + Z_i^T\gamma + \alpha_i + u_{it}$
- $X_{it}^T\beta$是可观测随时间变化的变量
- $Z_i^T\gamma$可观测不随时间变化的变量
- α_i不可观测不随时间变化的变量
- u_{it}不可观测随时间变化的变量
  - *应满足*：$E(\alpha_i|X_{it}, Z_i) \neq 0$
  - $E(u_{it}|X_{it}, Z_i, \alpha_i) = 0$

一、个体内差分估计法

- $Y_{it} = X_{it}^T\beta + Z_i^T\gamma + \alpha_i + u_{it}$
- 对每个个体变量取个体平均值：
  - $\overline{Y_i} = \frac{1}{T}\sum_{t=1}^{T} Y_{it}$　　$\overline{X_i} = \frac{1}{T}\sum_{t=1}^{T} X_{it}$
  - $\overline{Z_i} = \frac{1}{T}\sum_{t=1}^{T} Z_i$　　$\alpha i = \frac{1}{T}\sum_{t=1}^{T} \alpha_i$
  - $\overline{u_i} = \frac{1}{T}\sum_{t=1}^{T} u_{it}$
- $Y_{it} - \overline{Y_i} = (X_i t - \bar{X}_i)^T\beta + (z_i - z_i)^T\gamma - (\alpha_i - \alpha_i) + (u_{it} - \bar{u})$
- $Y_{it} - \bar{Y}_i = (X_{it} - \bar{X}_i)^T\beta + (u_{it} - \bar{u}_i)$
- $\widetilde{Y_{it}} = \tilde{X}_{it}^T\beta + \tilde{u}_{\mathrm{it}}$

一、个体内差分估计法（续）

- $\tilde{Y}_{it} = \tilde{X}_{it}^{T}\beta + \tilde{u}_{\mathrm{it}}$
- 运用OLS估计法可求得β：
- $\hat{\beta}^{i} = \left(\tilde{X}^{T}\tilde{X}\right)^{-1}\left(\tilde{X}^{T}\tilde{Y}\right) = \left(\sum_{i=1}^{N}\sum_{i=1}^{T}\tilde{X}_{it}\tilde{X}_{it}^{T}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{i=1}^{T}\tilde{X}_{it}\tilde{Y}_{it}\right) =$

$$\left(\sum_{\mathrm{i}=1}^{N}\sum_{t=1}^{T}(X_{it}-\bar{X}_{i})(X_{it}-\bar{X}_{i})^{T}\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T}(X_{it}-\bar{X}_{i})(Y_{it}-\bar{Y}_{i})\right)$$

- 这里的β称为个体内差分估计量、个体内估计量或固定效应估计量

二、最小二乘虚拟变量估计法(LSDV)

- $Y_{it} = X_{it}^T\beta + z_i^T\gamma + \alpha_1 D_1 + \cdots + \alpha_i D_i + \cdots + \alpha_N D_N + u_{it}$
- 这里，可观测且不随时间变化的变量Z_i与个体虚拟变量$(D_1 \ldots D_N)$会存在共线性，所以固定效应模型通常不包含Z_i，模型改写为：
- $Y_{it} = X_{it}^T\beta + \sum_{i=1}^{N} d_i D_i + u_{it}$

  ①$Y_{it} = \hat{\pi}_1 D_1 + \cdots + \hat{\pi}_i Di + \cdots + \hat{\pi}_N D_N + \hat{\varepsilon}_{it} = \hat{\pi}_i Di + \hat{\varepsilon}_{it}$
  - $x_{it} = \hat{\eta}_1 D_1 + \cdots + \hat{\eta}_i D_i + \cdots + \hat{\eta}_N D_N + \hat{\xi}_t = \hat{\eta}_i Di + \hat{\xi}_{it}$
  - $\hat{\varepsilon}_{it} = Y_{it} - \overline{Y_i} = \tilde{Y}_{it}$
  - $\hat{\xi}_{it} = x_{it} - \bar{x}_i = \tilde{x}_{it}$

  ②将残差值$\hat{\varepsilon}_{it}$对$\hat{\xi}_{it}$进行回归，得到回归系数即为所求β值(Frisch-Waugh-Lovell)
- LSDV方法缺点：N很大时引入虚拟变量过多，使得估计的计算量变得很大。实际运用中多使用个体内差分法

三、一阶差分估计法

- 对个体前后两期做差分以去除个体固定效应
- $Y_{it} = X_{\mathrm{it}}^{T}\beta + \mathrm{d}_i + u_{it}$ (1)
- $Y_{it-1} = X_{\mathrm{it-1}}^{T}\beta + \mathrm{d}_i + u_{it-1}$ (2)
- (1)式减去(2)式得:
- $\Delta Y_{it} = \Delta X_{\mathrm{it}}^{T}\beta + \Delta u_{it}$
- 由于ΔX_{it}与Δu_{it}不相关，故使用OLS方法得到的$\hat{\beta}^{FD}$是β的一致估计量

四、时间固定效应的引入

- 通过引入时间固定效应，可以控制同一时间不随个体变化的变量(如每年的宏观经济因素)，模型如下：
- 双向固定效应：$Y_{it} = X_{\mathrm{it}}^{T}\beta + \mathrm{d}_i + T_t + u_{it}$
- 通过引入T-1个时间虚拟变量控制年固定效应
- $Y_{it} = X_{\mathrm{it}}^{T}\beta + \mathrm{d}_i + \sum_{t=2}^{T}\gamma_t T_t + u_{it}$

五、个体效应估计残差

- 固定效应模型有：$Y_{it} = X_{it}^T\beta + \mathrm{d}_i + u_{it}$
- Z_i系数的估计
  - $\hat{d}_i = \bar{Y}_i - \bar{X}_i^T\hat{\beta}$
  - $\hat{d}_i$ 是 $\mathrm{d}_i = z_i^T\gamma + \alpha_i$的一致估计量
  - 当Z_i与α_i不相关时，通过$\hat{d}_i$对Z_i的回归可以求得Z_i的系数γ。但是通常Z_i与α_i相关，因此无法通过该方法估计γ
- “偶发参数问题”(Incidental Parameters Problem)
  导致估计的固定效应$\hat{d}_i$ 不一致,当时间T趋向无穷时，$\hat{d}_i$ 才是一致估计量

第六节面板数据分析实例

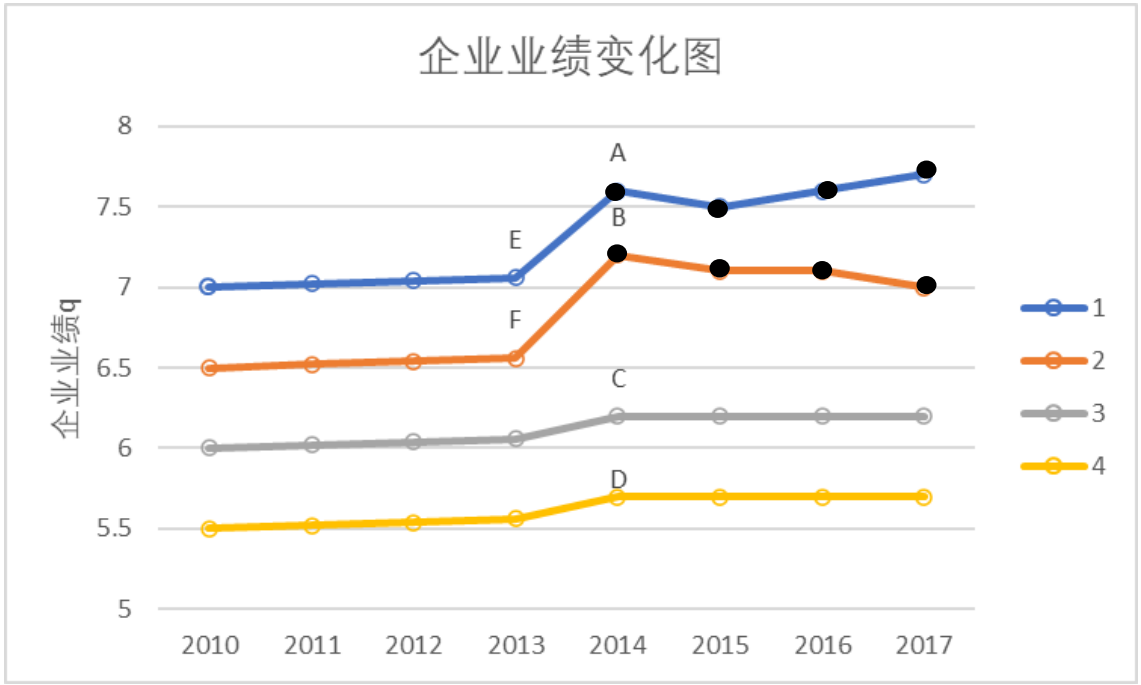
- A省在2014年进行了税法改革，B省没有进行税法改革。现抽取A、B两省各两个企业样本。数据包含id(企业代码)、year(年份)、q(企业业绩指标，越大越好)、tax(虚拟变量)

id	year	q	tax
1	2010	7	0
1	2011	7.02	0
1	2012	7.04	0
1	2013	7.06	0
1	2014	7.6	1
1	2015	7.5	1
1	2016	7.6	1
1	2017	7.7	1
2	2010	6.5	0
2	2011	6.52	0
2	2012	6.54	0
2	2013	6.56	0
2	2014	7.2	1
2	2015	7.1	1
2	2016	7.1	1
2	2017	7	1

id	year	q	tax
3	2010	6	0
3	2011	6.02	0
3	2012	6.04	0
3	2013	6.06	0
3	2014	6.2	0
3	2015	6.2	0
3	2016	6.2	0
3	2017	6.2	0
4	2010	5.5	0
4	2011	5.52	0
4	2012	5.54	0
4	2013	5.56	0
4	2014	5.7	0
4	2015	5.7	0
4	2016	5.7	0
4	2017	5.7	0

数据简单描述

- 受税改影响企业(1和2)在2014年后业绩较高的原因有：
  - 受税改影响企业业绩在2014年以前就较高
  - 2014年以后出现了提高企业业绩的其他宏观因素
  - 税改的结果



一、简单截面回归

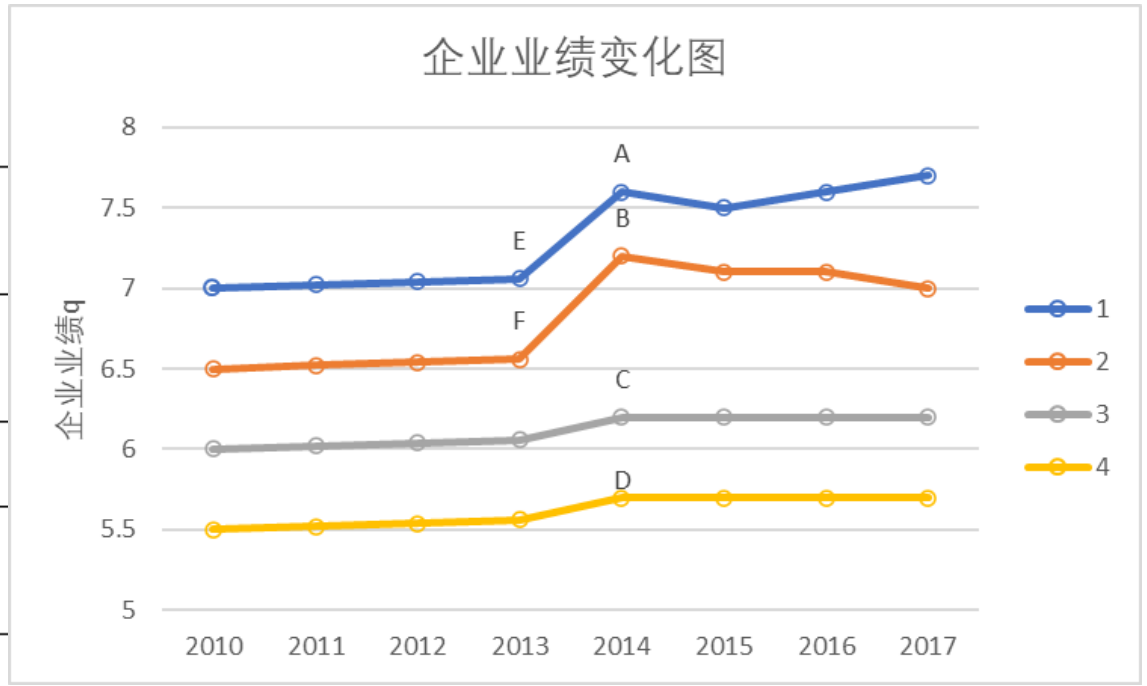
- $q_{i2014} = \alpha + \beta tax_{i2014} + u_{i2014}$
- 得到β=1.45=(A+B)/2-(C+D)/2

```
. tsset id year
        panel variable:  id (strongly balanced)
         time variable:  year, 2010 to 2017
                 delta:  1 unit
```

```
. reg q tax if year==2014
```

Source	SS
Model	2.1025
Residual	.205
Total	2.3075

q	Coef.
tax	1.45
_cons	5.95



二、合并横截面回归

- $q_{it} = \alpha + \beta tax_{it} + u_{it}, i = 1, \ldots, 4; t = 2010, \ldots, 2017$
- 对所有数据进行简单OLS回归，得到β=1.18

```
. reg q tax
```

Source	SS	df	MS	Number of obs	=	32
				F(1, 30)	=	37.81
Model	8.3544	1	8.3544	Prob > F	=	0.0000
Residual	6.6288	30	.22096	R-squared	=	0.5576
				Adj R-squared	=	0.5428
Total	14.9832	31	.483329032	Root MSE	=	.47006

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tax	1.18	.1919028	6.15	0.000	.7880823	1.571918
_cons	6.17	.0959514	64.30	0.000	5.974041	6.365959

三、固定效应模型LSDV估计

- $q_{it} = \beta tax_{it} + \alpha_1 firm_1 + \alpha_2 firm_2 + \alpha_3 firm_3 + \alpha_4 firm_4 + u_{it}$
  - 消除了受税改影响企业和不收税改影响企业在税改之前的差异，该β值包含了税改和2014年后宏观因素的影响

```
. tabulate id,gen(firm)

. reg q tax firm1-firm4, noconstant
```

Source	SS	df	MS	Number of obs	=	32
				F(5, 27)	=	44635.78
Model	1352.2988	5	270.45976	Prob > F	=	0.0000
Residual	.1636	27	.006059259	R-squared	=	0.9999
				Adj R-squared	=	0.9999
Total	1352.4624	32	42.26445	Root MSE	=	.07784

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tax	.57	.0389206	14.65	0.000	.4901415	.6498585
firm1	7.03	.0337062	208.57	0.000	6.96084	7.09916
firm2	6.53	.0337062	193.73	0.000	6.46084	6.59916
firm3	6.115	.027521	222.19	0.000	6.058531	6.171469
firm4	5.615	.027521	204.03	0.000	5.558531	5.671469

四、固定效应个体内估计

- $q_{it} = \beta tax_{it} + \alpha_i + u_{it}$， $i=1,\ldots,4;\ t=2010,\ldots,2017$

```
. egen mq=mean(q),by(id)

. egen mtax=mean(tax),by(id)

. generate within_q=q-mq

. generate within_tax=tax-mtax

. reg within_q within_tax,noconstant
```

Source	SS	df
Model	1.29959997	1
Residual	.163599998	31
Total	1.46319997	32

within_q	Coef.	Std. Err.
within_tax	.57	.0363229

```
. xtreg q tax,fe

Fixed-effects (within) regression               Number of obs      =        32
Group variable: id                              Number of groups   =         4

R-sq:                                           Obs per group:
     within  = 0.8882                                         min =         8
     between = 0.8521                                         avg =       8.0
     overall = 0.5576                                         max =         8

                                                F(1,27)            =    214.48
corr(u_i, Xb)  = 0.5066                         Prob > F           =    0.0000
```

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tax	.57	.0389206	14.65	0.000	.4901415	.6498585
_cons	6.3225	.0168531	375.15	0.000	6.28792	6.35708
sigma_u	.6020036					
sigma_e	.07784124					
rho	.98355552	(fraction of variance due to u_i)				

F test that all u_i=0: F(3, 27) = 355.67 Prob > F = 0.0000

五、一阶差分模型回归估计

- $\Delta q_{it} = \beta \Delta tax_{it} + u_{it}$， $i=1,\ldots,4;\ t=2010,\ldots,2017$

```
. generate dq=q-l.q
(4 missing values generated)

. generate dtax=tax-l.tax
(4 missing values generated)

. reg dq dtax, noconstant
```

Source	SS	df	MS		
Model	.696200008	1	.696200008	Prob > F	0.0000
Residual	.098999998	27	.003666667	R-squared	0.8755
				Adj R-squared	0.8709
Total	.795200006	28	.0284	Root MSE	.06055

Number of obs	=	28
F(1, 27)	=	189.87

dq	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
dtax	.59	.0428174	13.78	0.000	.5021459	.6778541

六、纳入时间固定效应

- $q_{it} = \beta tax_{it} + \alpha_i + \gamma_{2011} Year_{2011} + \gamma_{2012} Year_{2012} + \cdots + \gamma_{2017} Year_{2017} + u_{it}$
- 包含时间固定效应和企业固定效应

```
. xtreg q tax i.year, fe

Fixed-effects (within) regression               Number of obs      =        32
Group variable: id                              Number of groups   =         4

R-sq:                                           Obs per group:
     within  = 0.9761                                         min =         8
     between = 0.8521                                         avg =       8.0
     overall = 0.4280                                         max =         8

                                                F(8,20)            =    102.01
corr(u_i, Xb)  = 0.3827                         Prob > F           =    0.0000
```

q	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tax	.4	.0295804	13.52	0.000	.3382964	.4617036
year						
2011	.02	.0295804	0.68	0.507	-.0417036	.0817036
2012	.04	.0295804	1.35	0.191	-.0217036	.1017036
2013	.06	.0295804	2.03	0.056	-.0017036	.1217036
2014	.225	.0330719	6.80	0.000	.1560132	.2939868

第七节面板数据常见问题

- 一、是选择固定效应模型还是随机效应模型？
  - 取决于个体效应α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

  - Hausman检验判断使用RE模型还是FE模型
    - Hausman检验统计量：
    - $H = (\hat{\beta}^{FE} - \hat{\beta}^{RE})^T [\mathrm{Var}(\hat{\beta}^{FE}) - \mathrm{var}(\hat{\beta}^{RE})]^{-1} \times (\hat{\beta}^{FE} - \hat{\beta}^{RE}) \rightarrow \chi^2(k)$
    - K是模型中自变量的取值个数。若H接近于0，$\hat{\beta}^{FE}$与$\hat{\beta}^{RE}$一致，可使用RE模型，若H值较大，应选用FE模型
      - 多数情况下不需要Hausman检验，直接使用FE模型

第七节面板数据常见问题

- 二、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，应当如何理解这些变化？

①加入固定效应后，系数的大小和方向可能会发生变化

- 未使用固定效应模型时：$E(INC_{it}|EDU_{it}, GENDER_i) = (\alpha + \theta\phi_0) + (\beta + \theta\phi_1)EDU_{it} + (\gamma + \theta\phi_2)GENDER_i$
- 加入固定效应后：$E(INC_{it}|EDU_{it}, GENDER_i) = \alpha + \beta EDU_{it} + \gamma GENDER_i + \alpha_i$

②使用固定效应模型，通常估计系数方差会变大

- 个体间平均差异未被用来估计X和Y的关系，使用的信息变少了，造成估计系数方差变大

第七节面板数据常见问题

- 三、当使用固定效用模型后，有些变量系数变为不显著，是否就意味着该变量和被解释变量没有因果关系？

  ①变量X_{it}对Y_{it}没有因果影响

  ②变量X_{it}对Y_{it}有因果影响，但变量X_{it}的个体变化太小，控制固定效应后，由于信息不够造成估计系数方差太大而导致X_{it}系数不显著

  - 以教育对收入影响为例，$(EDU_{it} - \overline{EDU_i})$变化很小

- 运用固定效应模型需要变量有充分的个体内变化信息

谢谢！