PS1 解答*

赵之航
2018311178
国民经济管理 18

2020 年 10 月 16 日

1. （20 分） 考察一张 2013 至 2016 年的数据表中表示个人上报健康措施的变量 $\mathcal{H}$ ，探究一项自 2015 年起实施的医保政策对该变量的影响，已知政策覆盖不完全。分别定义两个虚拟变量

$$Policy_i = \begin{cases} 1, i\text{被医保政策覆盖} \\ 0, \text{其他} \end{cases}$$

以及

$$Post_t = \begin{cases} 1, \text{时间是 2015 或 2016 年} \\ 0, \text{其他} \end{cases}$$

回归模型 (1) 为

$$\mathcal{H}_{it} = \beta_0 + \beta_1 Policy_i \times Post_t + u_{it} \tag{1}$$

(a) 哪一组是参照组（即 β_0 表示哪一组)?

(b) 参数 β_1 的意义是什么?

一位研究者构造了一个更为复杂的饱和模型 (2)

$$\mathcal{H}_{it} = \beta_0 + \beta_1 Policy_i + \beta_2 Post_t + \beta_3 Policy_i \times Post_t + u_{it} \tag{2}$$

(c) $\beta_0, \beta_0+\beta_1, \beta_0+\beta_2, \beta_0+\beta_1+\beta_2+\beta_3$ 表示的实验组分别是哪个?

(d) 为什么不考察 $\beta_0+\beta_1+\beta_2$ 代表的实验组?

(e) β_3 是 $\beta_0, \beta_0+\beta_1, \beta_0+\beta_2, \beta_0+\beta_1+\beta_2+\beta_3$ 的线性组合吗? 如何理解 β_3 的意义?

解：

(a) 当 $Policy_i, Post_t$ 中至少有一个为零时，有 $\mathcal{H}_{it} = \beta_0$，故 β_0 表示未被医保政策覆盖或者 2013、2014 年上报的人群组成的参照组。

(b) β_1 表示 2013、2014 年上报的人群与医保覆盖范围两者对个人上报健康措施的统计差异。

(c) β_0 表示未被医保政策覆盖以及 2013、2014 年上报的人群
$\beta_0+\beta_1$ 表示被医保政策覆盖且在 2013、2014 年上报的人群
$\beta_0+\beta_2$ 表示未被医保政策覆盖且在 2015、2016 年上报的人群
$\beta_0+\beta_1+\beta_2+\beta_3$ 表示被医保政策覆盖且在 2015、2016 年上报的人群

(d) $\beta_0+\beta_1+\beta_2$ 没有意义，这由于 $\beta_0, \beta_1, \beta_2$ 均存在时，$Policy_i, Post_t$ 均为 1，故 β_3 必存在。

(e) 是。$\beta_3 = (\beta_0+\beta_1+\beta_2+\beta_3) - (\beta_0+\beta_1) - (\beta_0+\beta_2) + \beta_0$
β_3 表示 2013、2014 年上报的人群对个人上报健康措施统计结果的影响随医保覆盖范围的变化而变化。

*Powered by LaTeX

2. （20 分） 已知数据如表 (1)

Table 1: A Random Sample of Students' Critical Thinking Scores

Student ID	isID08	isID09	Year	isYr17	isYr18	DevPlan	Score
2016311407	0	0	2016	0	0	0	86
2016311408	1	0	2016	0	0	0	84
2016311409	0	1	2016	0	0	0	81
2016311407	0	0	2017	1	0	1	96
2016311408	1	0	2017	1	0	0	84
2016311409	0	1	2017	1	0	0	83
2016311407	0	0	2018	0	1	1	95
2016311408	1	0	2018	0	1	1	93
2016311409	0	1	2018	0	1	0	87

(a) 求解下列回归模型

$$Score_{it} = \beta_0 + \beta_1 DevPlan_{it} + \phi_i + \varphi_t + u_{it}, \tag{3}$$
$$Score_{it} = \gamma_0 + \pi_i + \lambda_t + v^y_{it}, \tag{4}$$
$$DevPlan_{it} = \theta_0 + \zeta_i + \xi_t + v^x_{it}, \tag{5}$$
$$\widehat{v}^y_{it} = \beta_1 \widehat{v}^x_{it} + v_{it} \tag{6}$$

其中 i 表示学生，t 表示年份，$\widehat{v}^y_{it} = Score_{it} - \widehat{\gamma_0} - \widehat{\pi_i} - \widehat{\lambda_t}$，$\widehat{v}^x_{it} = DevPlan_{it} - \widehat{\theta_0} - \widehat{\zeta_i} - \widehat{\xi_t}$

(b) 在回归表中增加一列以求解

$$\widehat{v}^y_{it} = \alpha_0 + \alpha_1 \widehat{v}^x_{it} + v_{it} \tag{7}$$

$\widehat{\alpha}_1$ 与 $\widehat{\beta}_1$ 接近吗？为什么？

解：

(a) 如下图

```
. reg score devplan isid08 isid09 isyr17 isyr18
```

Source	SS	df	MS	Number of obs	=	9
				F(5, 3)	=	15.63
Model	238.833333	5	47.7666667	Prob > F	=	0.0233
Residual	9.16666667	3	3.05555556	R-squared	=	0.9630
				Adj R-squared	=	0.9014
Total	248	8	31	Root MSE	=	1.748

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
devplan	6.5	2.140872	3.04	0.056	-.3132105	13.31321
isid08	-3.166667	1.595712	-1.98	0.141	-8.244934	1.911601
isid09	-4.333333	2.018434	-2.15	0.121	-10.75689	2.090223
isyr17	1.833333	1.595712	1.15	0.334	-3.244934	6.911601
isyr18	3.666667	2.018434	1.82	0.167	-2.75689	10.09022
_cons	86.16667	1.485527	58.00	0.000	81.43906	90.89428

Figure 1: 模型 (3)

```
. reg score isid08 isid09 isyr17 isyr18
```

Source	SS	df	MS	Number of obs	=	9
				F(4, 4)	=	5.64
Model	210.666667	4	52.6666667	Prob > F	=	0.0612
Residual	37.3333333	4	9.33333333	R-squared	=	0.8495
				Adj R-squared	=	0.6989
Total	248	8	31	Root MSE	=	3.0551

score	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
isid08	-5.333333	2.494438	-2.14	0.099	-12.259	1.592338
isid09	-8.666667	2.494438	-3.47	0.025	-15.59234	-1.740996
isyr17	4	2.494438	1.60	0.184	-2.925671	10.92567
isyr18	8	2.494438	3.21	0.033	1.074329	14.92567
_cons	88.33333	2.2771	38.79	0.000	82.01109	94.65558

```
. predict vity, res
```

Figure 2: 模型 (4)

```
. reg devplan isid08 isid09 isyr17 isyr18
```

Source	SS	df	MS	Number of obs	=	9
				F(4, 4)	=	2.00
Model	1.33333333	4	.333333333	Prob > F	=	0.2593
Residual	.666666667	4	.166666667	R-squared	=	0.6667
				Adj R-squared	=	0.3333
Total	2	8	.25	Root MSE	=	.40825

devplan	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
isid08	-.3333333	.3333333	-1.00	0.374	-1.258815	.5921484
isid09	-.6666667	.3333333	-2.00	0.116	-1.592148	.258815
isyr17	.3333333	.3333333	1.00	0.374	-.5921484	1.258815
isyr18	.6666667	.3333333	2.00	0.116	-.258815	1.592148
_cons	.3333333	.3042903	1.10	0.335	-.511512	1.178179

```
. predict vitx, res
```

Figure 3: 模型 (5)

```
. reg vity vitx, noconstant
```

Source	SS	df	MS	Number of obs	=	9
				F(1, 8)	=	24.58
Model	28.1666669	1	28.1666669	Prob > F	=	0.0011
Residual	9.16666689	8	1.14583336	R-squared	=	0.7545
				Adj R-squared	=	0.7238
Total	37.3333338	9	4.1481482	Root MSE	=	1.0704

vity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vitx	6.5	1.311011	4.96	0.001	3.476803	9.523197

Figure 4: 模型 (6)

(b) 增加一列后的回归结果如下

```
. reg vity vitx
```

Source	SS	df	MS	Number of obs	=	9
				F(1, 7)	=	21.51
Model	28.1666669	1	28.1666669	Prob > F	=	0.0024
Residual	9.16666689	7	1.30952384	R-squared	=	0.7545
				Adj R-squared	=	0.7194
Total	37.3333338	8	4.66666673	Root MSE	=	1.1443

vity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
vitx	6.5	1.40153	4.64	0.002	3.185909	9.814091
_cons	1.32e-08	.3814481	0.00	1.000	-.9019814	.9019814

Figure 5: 模型 (7)

$\widehat{\alpha}_1$ 与 $\widehat{\beta}_1$ 接近，这由于截距项的 t 值为 0.00，$P > |t|$ 的概率为 1.000，截距项必趋近于零。

3. （40 分） 阅读参考文献，主要模型为

$$Y_{igast} = \alpha_0 + \alpha_1 Eligible_g + \alpha_2 (Eligible_g \times Post_t) + \beta X_{ig} + \gamma_{st} + \gamma_{rt} + \gamma_{at} + \epsilon_{igast} \tag{8}$$

(a) 为什么不加入个体固定效应（如 γ_i）?

(b) 为什么不加入年份与资格固定效应（如 γ_{gt}）?

(c) 为什么不加入 $\alpha_3 Post_t$?

(d) 你认为 $Eligible_g$ 的系数 α_1 可估计吗？请说明理由。

(e) 在加入**移民与资格的年龄，年龄与资格固定效应**以及**年份固定效应**后，作者需要加入**移民年份**这一虚拟变量吗？若需要，给出该虚拟变量可估计的实例；若不需要，给出理由。

(f) 你能给出不在现有控制变量或固定效应之内的其他控制变量或固定效应吗？你认为这是必要的吗?

解：

(a) 个体需要固定的要素过多，包含移民时间、出生地区、移民年龄等等，方便起见，文中使用了 X_{ig} 这个控制变量，故再添加个体固定效应是不必要的。

(b) 年份与资格的固定效应对不同的个体都会不同，这由于同一年份不同个体的状态不同，真正有意义的是年龄。文中加入了年龄与资格固定效应。

(c) 加入 $Post_t$ 会影响估计的准确性，这由于政策的效应具有滞后性，个体对政策的反应需要时间。政策刚施行时，个体来不及对现有的计划做出相应的调整，如果加入会弱化政策在模型中的作用，与实际情况产生偏差。

(d) 我认为可估计，这由于交互项 $Eligible_g \times Post_t$ 包含了政策对个体资格的影响，排除了干扰。

(e) 不需要。根据已有的变量可以推算出个体的移民年份，再加入这一虚拟变量会导致多重共线性。

(f) 加入新的控制变量——社区平均犯罪率，这是必要的，社区的治安情况将极大地影响青少年的发展。

4. （20 分） 在参考文献中，作者将主要模型中的标准误按州进行了聚类。
   (a) 请举出两个观测值误差项具有相关性的实例。
   (b) 请举出两个观测值误差项不具有相关性的实例。
   (c) 有的研究者可能希望按州和年份进行聚类，这就是所谓的” 双向聚类标准误”。请举出两个观测值误差项不具有相关性的实例。
   (d) 你更喜欢哪种聚类标准误，是以州聚类的，还是以州和年聚类的?

解：

(a) 同一社区的两名移民可能会相互影响，如教唆吸毒、帮派犯罪等。

(b) 富人社区与平民社区的移民相互影响的可能性较低。

(c) 两名移民所在州的社区、移民年份（涉及到移民时的政策）以及原籍不同，他们的误差项不具有相关性。

(d) 我更偏向以州和年聚类的，因为这种方式可以更好地排除内生因素的干扰，尽可能使模型的影响因素外生。