附录

Part Ⅰ

为了便于理解本文准标准化转换的改进思路，这里以 Abadie 等（2010）所研究的加州禁烟法案问题为例进行说明。

子图 A1(a) 和子图 1(b) 呈现的是 Abadie 等（2010）安慰剂检验中，加州（子图 A1(a)）和罗德岛州（图 A1(b)）使用合成控制法估计的结果。在干预前，如果以加州拟合为标准，那么罗德岛州的“反事实”估计结果并未很好地拟合罗德岛州真实的香烟消费量，因此，从因果推断的角度来看，本文并不认为罗德岛州干预后时期的禁烟效果完全来自于禁烟法案的实施，可能还存在噪音干预的成分。所以在图 A1(d) 的安慰剂检验中，尽管干预后罗德岛州显示的虚拟的禁烟效果高于加州，但是考虑到罗德岛州干预前的合成效果不佳的问题，因此该结论仍然饱受质疑。这就是政策干预不可比问题。

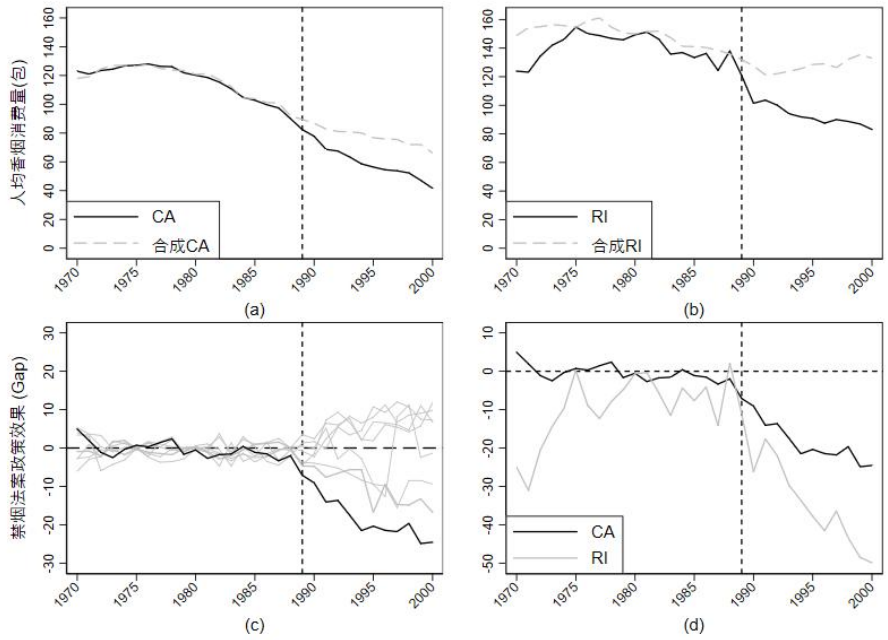

图 A1 禁烟法案案例分析

注：图中 CA 表示加州，RI 表示罗德岛州。子图 A1(a) 和 A1(b) 分别呈现的是 Abadie 等（2010）中加州和罗德岛州禁烟法案合成控制估计结果，子图 A1(d) 加州和罗德岛州安慰剂检验结果，子图 A1(c) 是删除大于加州 1.5 倍以上州后的安慰剂检验结果，黑色实线代表加州，灰色实线代表其他州。纵向虚线代表禁烟法案发生时间 (1989 年)，横向虚线代表无政策效果，即 $Gap_{jt}=0$。

为此，Abadie 等（2010）建议删除干预前（1989 年之前）时段内合成效果不佳的控制组样本，以便尽可能控制噪音成分对 p 值估计结果的影响。然而，困难在于——如何界定合成效果不佳？Abadie 等（2010）的思路是：以加州（实验组）在干预前时段内的预测均方误差（mean square prediction error，*MSPE*）为判断基准，记为 $MSPE_{CA}$，进而设定一

个临界倍数 k （取值为 20, 10, 5, 2 等）。若 $MSPE_j > k \times MSPE_{CA}$，则认为其噪音过大，并予以删除（$MSPE_j$ 表示控制组中第 j 个州的 $MSPE$）。最终的经验 p 值是基于删减处理后得到的相对“干净”的样本计算而得。

子图 A1(c) 是选取 $k=1.5$ 得到的安慰剂检验结果。从干预前时期的拟合效果来看，多数控制组州的合成效果都比较好，意味着噪音成分大幅下降了。从干预后时期来看，加州的政策效果最为明显（黑色实线在干预后时期都位于最下方）。然而，当我们再次计算经验 p 值时就面临一个非常尴尬的问题：由于这个“干净”的样本中只包含了 7 个控制组 ①，经验 p 值 $=1/8=0.125$。按照常规的评判标准（p 值小于 0.10 或 0.05），这意味着加州的政策效应并不具有统计上的显著性。

由于 k 的选择取决于研究者的偏好，那么聪明的研究者通过选择一个适当 k 值，以便让经验 p 值小于 0.10 / 0.05（10% / 5% 水平显著）。这就是正文中提及的对显著性结果“主观挑选”（cherry-picking）问题。

Part Ⅱ

为了说明本文方法的适用性，这里以合成控制法领域中的另一篇备受关注的文献为例来展示本文准标准化修正方法的有效性。在 Abadie 等（2015）的研究中，他们分析了两德统一这一事件对于西德人均 GDP 的影响。相比于 Abadie 等（2010），该文中的控制组仅包含 17 个国家，样本量更小了。子图 A2(a) 源于 Abadie 等（2015）文中的图 5，展示了包含所有控制组国家的安慰剂检验结果。显然，在干预后时段内的任何一个年份上，都至少有 3 个控制组国家的政策效果强于西德，意味着对应的经验 p 值 $=3/17=0.176$。然而，这一结果并不足信，因为干预前时期（1960—1990 年）有多个控制组国家的拟合效果欠佳，包含严重的噪音成分。为此，Abadie 等（2015）转而使用预测均方根误差的比值作为安慰剂检验的统计量来进行检验。

在子图 A2(b) 中，本文沿用 Abadie 等（2010）的作法，仅保留控制组国家预测均方误差（$MSPE_i$）小于西德 20 倍的国家。但是出现了尴尬的结果，此时控制组国家仅有 4 个，经验 p 值 $=1/5=0.2$，在传统的显著性判断标准条件下，我们认为两德统一并未减少西德人均 GDP 水平。

为此本文使用准标准化转换方法进行修正子图 A2(b) 安慰剂检验过程，子图 A2(c) 中显示，从 1993 开始，西德人均 GDP 显著低于控制组国家，其显著性水平 p 值 $=1/17=0.059$，这一结果与 Abadie 等（2015）图 5 结果一致（如子图 A2(d)）②。但是相比于 Abadie 等（2015）的图 5 结果，子图 A2(c) 动态展现了干预后时期（1991—2003）政策效果的变化情况，能够更好地分析两德统一对于西德人均 GDP 影响的滞后性。

---

① 与 Abadie 等（2010）不同的是，由于安慰剂检验中使用嵌套法计算最有权重的方法存在部分控制组无法收敛的现象。所以为了进行比较，本文不使用嵌套法，因此权重估计结果与 Abadie 等（2010）结果存在细微的差异现象，但是这种现象并不影响结论的分析。当删除均方误差根大于加州 1.5 倍的控制组后，剩余的样本数量为 8 个，分别为：California (3)、Georgia (7)、Louisiana (14)、Misouri (18)、Montana (19)、Nebraska (20)、New Mexico (23)、South Carolina (30)。

② 由于使用交叉验证法进行估计时，存在部分国家不收敛问题，因此子图 A2 并未使用嵌套方法进行估计，因而使得子图 A2(d) 与 Abadie 等（2015）的图 5 存在一定的差异，但是显著性结论是一致的。

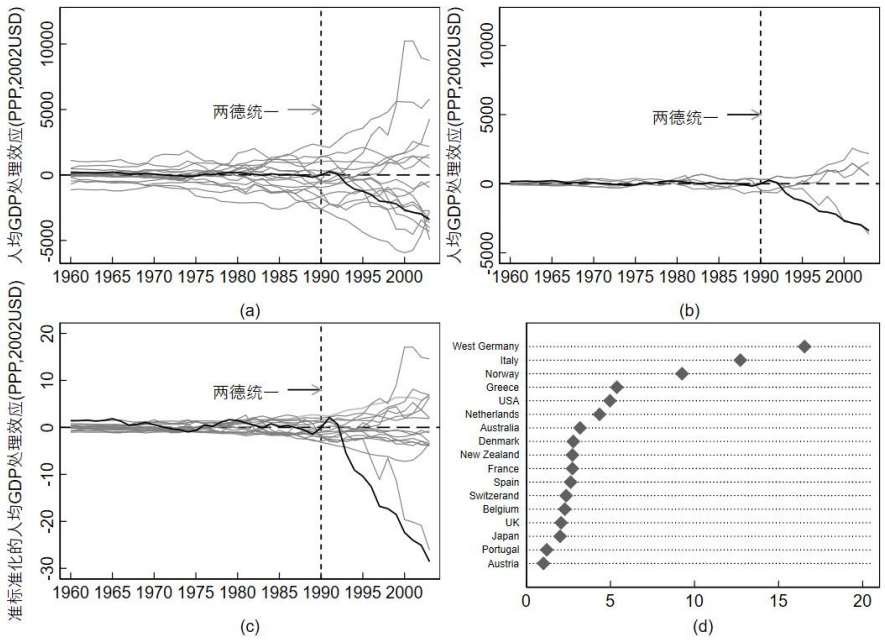

图 A2 两德统一对西德经济的影响的安慰剂检验

注：子图 A2(a) 是使用 Abadie 等 (2010) 安慰剂检验方法估计两德统一对于西德人均 GDP 的影响，子图 A2(b) 是删除大于西德 20 倍均方误差的国家后的安慰剂检验结果，子图 A2(c) 使用准标准化修正后的安慰剂检验结果，子图 A2(d) 是使用 Abadie 等 (2015) 方法，利用干预后与干预前预测均方根误差之比作为安慰剂检验统计量。子图 A2(a)、A2(b) 和 A2(c) 中的纵向虚线代表两德统一时间 (1990 年)，横向虚线代表无政策效果，即 Gap_{jt}=0，其中 j=1, 2, ⋯, 17。黑色实线代表西德，灰色实线代表其他国家。

Part Ⅲ

表 A1 所描述的是通过 Bootstrap 方式构建的置信水平为 95%和 99%的置信区间。其中，列 (1) 描述的是加州在整个研究期间的标准化后的政策效果 ($Gap^{S}_{CA,t}$)。第 (2) ～(3) 列描述了经验分布和正态分布条件下的置信区间。

列 (2) 中，干预后时期 (1989—2000 年) 加州香烟消费量的变化均在 5%的显著性水平下显著，并且随时间显著减少，说明禁烟法案减少了香烟消费量。这一结果与正文中的子图 1(f) 与子图 1(e) 结果一致，说明我们的置信区间估计结果很好地揭示了加州禁烟法案对于香烟消费量减少的影响。

与干预后时期相比，干预前时期 (1970—1988 年) 中，除了 1988 年由于存在政策迁移导致加州香烟消费量在 5%水平上显著外，1975 年和 1981 年显著可能是由于特定时点的冲击造成的结果，对此为了能够很好地说明其对于禁烟法案政策效果不存在叠加效果，我们需要针对其进行时间安慰剂检验，或者通过分析该事件的随机性，以保证干预后政策效果的稳健性。但是对于 1970 年的香烟的显著性，由于缺少更长的时间跨度，因而无法进行时间安慰剂检验。但是结合 1971—1987 年后绝大部分时点政策效果的不显著的特点，我们也能够近似认为这是由于随机事件冲击导致的结果。

列 (1) 和列 (3) 的结果是基于正态分布假设下的估计结果。与列 (2) 结论一致，干预后时期的政策效果显著不为零，并且其在 0.1%的显著性水平下显著。但是在干预前时期，

进一步增加了 1971 年、1978 年、1979 年、1982 年、1986 年和 1987 年结果在 1%的显著性水平下显著非零。这一结果可能是由于抽样次数的较少，导致该时点的经验分布与正态分布之间存在较大的差异，因而随着抽样次数的增长，这种不一致问题将进一步得到改善。

表 A1　　基于自抽样（Bootstrap）构造的置信区间

时间	禁烟效果（$Gap^{S}_{CA,t}$）	Bootstrap 500 次置信区间（95%）	
	（1）	（2）经验分布置信区间	（3）正态分布置信区间
1970	2.433***	[1.253, 3.378]	[1.394, 3.471]
1971	0. 949*	[−0.000, 1.893]	[0.087, 1.811]
1972	–0.555	[−1.292, 1.511]	[−2.009, 0.899]
1973	–1.226	[−1.550, 1.159]	[−2.829, 0.376]
1974	–0.163	[−0.298, 1.055]	[−0.908, 0.582]
1975	0.350	[0.065, 0.923]	[−0.087, 0.787]
1976	0.153	[−0.669, 0.755]	[−0.452, 0.757]
1977	0.705	[−0.758, 0.786]	[−0.160, 1.570]
1978	1.173**	[−0.479, 1.222]	[0.236, 2.110]
1979	–0.807**	[−0.973, 0.110]	[−1.404, −0.210]
1980	–0.256	[−0.799, 0.083]	[−0.715, 0.202]
1981	–1.320**	[−1.675, −0.156]	[−2.195, −0.446]
1982	–0.846*	[−1.679, 0.117]	[−1.763, 0.071]
1983	–0.765	[−1.685, 0.312]	[−1.832, 0.302]
1984	0.196	[−1.800, 1.338]	[−1.659, 2.051]
1985	–0.558	[−1.366, 0.126]	[−1.236, 0.120]
1986	–0.750**	[−1.322, 0.004]	[−1.395, −0.104]
1987	–1.640**	[−1.916, 0.107]	[−2.771, −0.508]
1988	–0.976*	[−1.744, −0.020]	[−1.855, −0.097]
1989	–3.481***	[−4.231, −0.700]	[−5.541, −1.422]
1990	–4.452***	[−4.622, −1.494]	[−6.403, −2.501]
1991	–6.907***	[−6.939, −3.363]	[−8.961, −4.853]
1992	–6.701***	[−6.743, −3.504]	[−8.593, −4.809]
1993	–8.586***	[−8.664, −3.996]	[−11.245, −5.927]
1994	–10.551***	[−10.646, −4.712]	[−13.897, −7.205]
1995	–9.995***	[−10.073, −5.263]	[−13.008, −6.981]
1996	–10.526***	[−10.623, −5.096]	[−13.982, −7.070]
1997	–10.695***	[−10.883, −5.058]	[−14.432, −6.959]
1998	–9.640***	[−9.730, −5.278]	[−12.376, −6.905]
1999	–12.208***	[−12.315, −5.683]	[−16.221, −8.194]
2000	–12.042***	[−12.159, −5.572]	[−16.009, −8.074]

注：* 5% ** 1% ***0.1%，且列（1）的显著性水平是基于列（3）置信区间进行判断的。干预前时期为 1970—1988 年，干预后时期为 1989—2000 年。第 (1) 列的禁烟效果是经过准标准化转换后的结果。列（2）置信区间是根据分位数方法构建的结果，而列（3）则根据正态分布设定临界值进行构造的结果。

图 A3 是表 A1 中的列（3）的可视化结果。

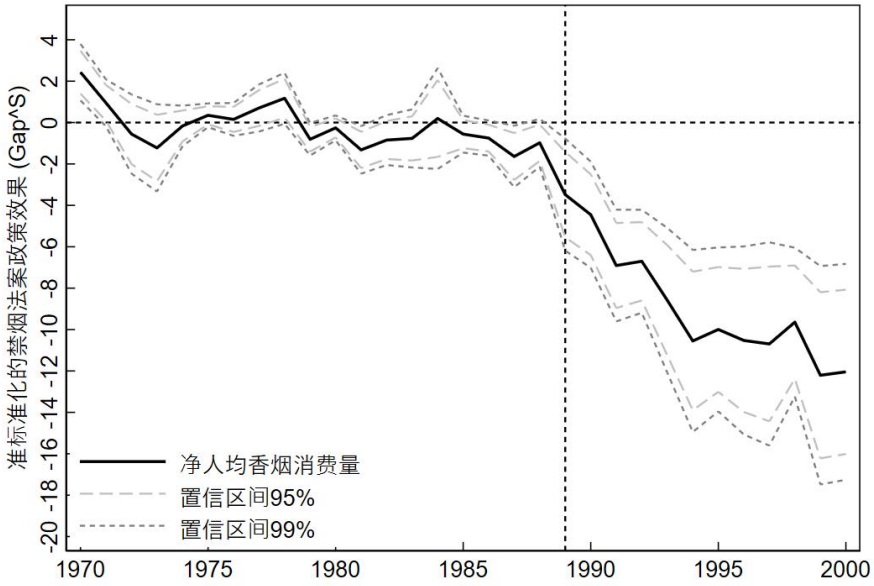

图 A3 禁烟法案置信区间

注：图中纵向虚线代表禁烟法案发生时间（1989 年），横向虚线代表无政策效果，即 $Gap^{S}_{CA,t}=0$ 。

参考文献

[1] Abadie A，Diamond A，Hainmueller J. Synthetic Control Methods for Comparative Case Studies：Estimating the Effect of 加州's Tobacco Control Program [J]. Journal of the American Statistical Association，2010，105(490)：493-505.

[2] Abadie A，Diamond A，Hainmueller J. Comparative Politics and the Synthetic Control Method [J]. American Journal of Political Science，2015，59(2)：495-510.